Prelims Probability

Christina Goldschmidt

Michaelmas Term 2012
(This version: 4th December 2012)

Background

Probability theory is one of the fastest growing areas of mathematics. Probabilistic arguments are used in an incredible range of applications from number theory to genetics, from geometry to finance. It is a core part of computer science and a key tool in analysis. And of course it underpins statistics. It is a subject that impinges on our daily lives: we come across it when we go to the doctor or buy a lottery ticket, but we're also using probability when we listen to the radio or use a mobile phone, or when we enhance digital images and when our immune system fights a cold. Whether you knew it or not, from the moment you were conceived, probability theory played an important role in your life.

We all have some idea of what probability is: maybe we think of it as an approximation to long run frequencies in a sequence of repeated trials, or perhaps as a measure of degree of belief warranted by some evidence. Each of these interpretations is valuable in certain situations. For example, the probability that I get a head if I flip a coin is sensibly interpreted as the proportion of heads I get if I flip that same coin many times. But there are some situations where it simply does not make sense to think of repeating the experiment many times. For example, the probability that 'UK interest rates will be more than 6% next March' or the probability that 'I'll be involved in a car accident in the next twelve months' cannot be determined by repeating the experiment many times and looking for a long run frequency.

The philosophical issue of interpretation is not one that we'll resolve in this course. What we *will* do is set up the abstract framework necessary to deal with complicated probabilistic questions.

These notes are intended to complement the contents of the lectures. They contain more material than the lectures and, in particular, a few more examples. You are nonetheless **strongly encouraged** to attend all of the lectures. An original version of the first part of these notes was written by Alison Etheridge and I have also made extensive use of notes by Neil Laws and Jonathan Marchini. I am very grateful to them all. The responsibility for any errors or inaccuracies is mine. Please send any comments or corrections to `goldschm@stats.ox.ac.uk`.

The synopsis and reading list from the course handbook are reproduced on the next page for your convenience. The suggested texts are an excellent source of further examples.

I hope you enjoy the course!

Overview

An understanding of random phenomena is becoming increasingly important in today's world within social and political sciences, finance, life sciences and many other fields. The aim of this introduction to probability is to develop the concept of chance in a mathematical framework. Random variables are introduced, with examples involving most of the common distributions.

Learning Outcomes

Students should have a knowledge and understanding of basic probability concepts, including conditional probability. They should know what is meant by a random variable, and have met the common distributions and their probability mass functions. They should understand the concepts of expectation and variance of a random variable. A key concept is that of independence which will be introduced for events and random variables.

Synopsis

Motivation, relative frequency, chance. Sample space, algebra of events, probability measure. Permutations and combinations, sampling with or without replacement. Conditional probability, partitions of the sample space, theorem of total probability, Bayes' Theorem. Independence.

Discrete random variables, probability mass functions, examples: Bernoulli, binomial, Poisson, geometric. Expectation: mean and variance. Joint distributions of several discrete random variables. Marginal and conditional distributions. Independence. Conditional expectation, theorem of total probability for expectations. Expectations of functions of more than one discrete random variable, covariance, variance of a sum of dependent discrete random variables.

Solution of first and second order linear difference equations. Random walks (finite state space only).

Probability generating functions, use in calculating expectations. Random sample, sums of independent random variables, random sums. Chebyshev's inequality, Weak Law of Large Numbers.

Continuous random variables, cumulative distribution functions, probability density functions, examples: uniform, exponential, gamma, normal. Expectation: mean and variance. Functions of a single continuous random variable. Joint probability density functions of several continuous random variables (rectangular regions only). Marginal distributions. Independence. Expectations of functions of jointly continuous random variables, covariance, variance of a sum of dependent jointly continuous random variables.

Textbooks

1. G. R. Grimmett and D. J. A. Welsh, *Probability: An Introduction*, Oxford University Press, 1986, Chapters 1–4, 5.1–5.4, 5.6, 6.1, 6.2, 6.3 (parts of), 7.1–7.3, 10.4.
2. J. Pitman, *Probability*, Springer-Verlag, 1993.
3. S. Ross, *A First Course In Probability*, Prentice-Hall, 1994.
4. D. Stirzaker, *Elementary Probability*, Cambridge University Press, 1994, Chapters 1–4, 5.1–5.6, 6.1–6.3, 7.1, 7.2, 7.4, 8.1, 8.3, 8.5 (excluding the joint generating function).

Chapter 1

Events and probability

1.1 Introduction

We will think of performing an experiment which has a set of possible outcomes Ω. We call Ω the *sample space* and its elements *sample points*. For example,

(a) tossing a coin: $\Omega = \{H, T\}$;

(b) throwing two dice: $\Omega = \{(i, j) : 1 \leq i, j \leq 6\}$.

A subset of Ω is called an *event*. An event $A \subseteq \Omega$ *occurs* if, when the experiment is performed, the outcome $\omega \in \Omega$ satisfies $\omega \in A$. You should think of events as things you can decide have or have not happened by looking at the outcome of your experiment. For example,

(a) coming up heads: $A = \{H\}$;

(b) getting a total of 4: $A = \{(1, 3), (2, 2), (3, 1)\}$.

The complement of A is $A^c := \Omega \setminus A$ and means "A does not occur". For events A and B,

$A \cup B$ means "at least one of A and B occurs";

$A \cap B$ means "both A and B occur";

$A \setminus B$ means "A occurs but B does not".

If $A \cap B = \emptyset$ we say that A and B are *disjoint* – they cannot both occur.

We assign a *probability* $\mathbb{P}(A) \in [0, 1]$ to each (suitable) event. For example,

(a) for a fair coin, $\mathbb{P}(A) = 1/2$;

(b) for two unweighted dice, $\mathbb{P}(A) = 1/12$.

(b) demonstrates the importance of *counting* in the situation where we have a finite number of possible outcomes to our experiment, all equally likely. For (b), Ω has 36 elements (6 ways of choosing i and 6 ways of choosing j). Since $A = \{(1,3),(2,2),(3,1)\}$ contains 3 sample points, and all sample points are equally likely, we get $\mathbb{P}(A) = 3/36 = 1/12$.

We want to be able to tackle much more complicated counting problems.

1.2 Counting

Most of you will have seen this before. If you haven't, or if you find it confusing, then you can find more details in the first chapter of *Introduction to Probability* by Ross.

Arranging distinguishable objects

Suppose that we have n distinguishable objects (e.g. the numbers $1, 2, \ldots, n$). How many ways to order them (*permutations*) are there? If we have three objects a, b, c then the answer is 6: *abc*, *acb*, *bac*, *bca*, *cab* and *cba*.

In general, there are n choices for the first object in our ordering. Then, whatever the first object was, we have $n-1$ choices for the second object. Carrying on, we have $n-m+1$ choices for the mth object and, finally, a single choice for the nth. So there are

$$n(n-1)\ldots 2.1 = n!$$

different orderings.

Since $n!$ increases extremely fast, it is sometimes useful to know Stirling's formula:

$$n! \sim \sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n},$$

where $f(n) \sim g(n)$ means $f(n)/g(n) \to 1$ as $n \to \infty$. This is astonishingly accurate even for quite small n. For example, the error is of the order of 1% when $n = 10$.

Arrangements when not all objects are distinguishable

What happens if not all the objects are distinguishable? For example, how many different arrangements are there of a, a, a, b, c, d?

If we had a_1, a_2, a_3, b, c, d, there would be 6! arrangements. Each arrangement (e.g. b, a_2, d, a_3, a_1, c) is one of 3! which differ only in the ordering of a_1, a_2, a_3. So the 6! arrangements fall into groups of size 3! which are indistinguishable when we put $a_1 = a_2 = a_3$. We want the number of groups which is just 6!/3!.

We can immediately generalise this. For example, to count the arrangements of a, a, a, b, b, d, play the same game. We know how many arrangements there are if the b's are distinguishable, but then all such arrangements fall into pairs which differ only in the ordering of b_1, b_2, and we see that the number of arrangements is 6!/3!2!.

Lemma 1.1. *The number of arrangements of the n objects*

$$\underbrace{\alpha_1, \ldots, \alpha_1}_{m_1 \text{ times}}, \underbrace{\alpha_2, \ldots, \alpha_2}_{m_2 \text{ times}}, \ldots, \underbrace{\alpha_k, \ldots, \alpha_k}_{m_k \text{ times}}$$

where α_i appears m_i times and $m_1 + \cdots + m_k = n$ is

$$\frac{n!}{m_1! m_2! \cdots m_k!}. \tag{1.1}$$

Example 1.2. *The number of arrangements of the letters of STATISTICS is $\frac{10!}{3!3!2!}$.*

If there are just two types of object then, since $m_1 + m_2 = n$, the expression (1.1) is just a binomial coefficient, $\binom{n}{m_1} = \frac{n!}{m_1!(n-m_1)!} = \binom{n}{m_2}$.

Note: we will always use the notation

$$\binom{n}{m} = \frac{n!}{m!(n-m)!}.$$

Recall the binomial theorem,

$$(x+y)^n = \sum_{m=0}^{n} \binom{n}{m} x^m y^{n-m}.$$

You can see where the binomial coefficient comes from because writing

$$(x+y)^n = (x+y)(x+y)\cdots(x+y)$$

and multiplying out, each term involves one pick from each bracket. The coefficient of $x^m y^{n-m}$ is the number of sequences of picks that give x exactly m times and y exactly $n-m$ times and that's the number of ways of choosing the m "slots" for the x's.

The expression (1.1) is called a *multinomial* coefficient because it is the coefficient of $a_1^{m_1} \cdots a_n^{m_k}$ in the expansion of

$$(a_1 + \cdots + a_k)^n$$

where $m_1 + \cdots m_k = n$. We sometimes write

$$\binom{n}{m_1 \; m_2 \; \ldots \; m_k}$$

for the multinomial coefficient.

Instead of thinking in terms of arrangements, we can think of our binomial coefficient in terms of choices. For example, if I have to choose a committee of size k from n people, there are $\binom{n}{k}$ ways to do it. To see how this ties in, stand the n people in a line. For each arrangement of k 1's and $n-k$ 0's I can create a different committee by picking the ith person for the committee if the ith term in the arrangement is a 1.

Many counting problems can be solved by finding a bijection (that is, a one-to-one correspondence) between the objects we want to enumerate and other objects that we already know how to enumerate.

Example 1.3. *How many distinct non-negative integer-valued solutions of the equation*

$$x_1 + x_2 + \cdots + x_m = n$$

are there?

Solution. Consider a sequence of n $\star$'s and $m-1$ $|$'s. There is a bijection between such sequences and non-negative integer-valued solutions to the equation. For example, if $m=4$ and $n=3$,

$$\underbrace{\star\,\star}_{x_1=2}\big|\underbrace{}_{x_2=0}\big|\underbrace{\star}_{x_3=1}\big|\underbrace{}_{x_4=0}$$

There are $\binom{n+m-1}{n}$ sequences of n $\star$'s and $m-1$ $|$'s and, hence, the same number of solutions to the equation. □

It is often possible to perform quite complex counting arguments by manipulating binomial coefficients. Conversely, sometimes one wants to prove relationships between binomial coefficients and this can most easily be done by a counting argument. Here is one famous example:

Lemma 1.4 (Vandermonde's identity)**.** *For $k, m, n \geq 0$,*

$$\binom{m+n}{k} = \sum_{j=0}^{k} \binom{m}{j}\binom{n}{k-j}, \tag{1.2}$$

where we use the convention $\binom{m}{j} = 0$ for $j > m$.

Proof. Suppose we choose a committee consisting of k people from a group of m men and n women. There are $\binom{m+n}{k}$ ways of doing this which is the left-hand side of (1.2).

Now the number of men in the committee is some $j \in \{0, 1, \ldots, k\}$ and then it contains $k-j$ women. The number of ways of choosing the j men is $\binom{m}{j}$ and for each such choice there are $\binom{n}{k-j}$ choices for the women who make up the rest of the committee. So there are $\binom{m}{j}\binom{n}{k-j}$ committees with exactly j men and summing over j we get that the total number of committees is given by the right-hand side of (1.2). □

"Breaking things down" is an important technique in counting - and also, as we'll see, in probability.

1.3 The axiomatic approach

Definition 1.5. *A* probability space *is a triple $(\Omega, \mathcal{F}, \mathbb{P})$ where*

1. *Ω is the* sample space*,*
2. *$\mathcal{F}$ is a collection of subsets of Ω, called* events*, satisfying axioms $\boldsymbol{F}_1$–$\boldsymbol{F}_3$ below,*
3. *$\mathbb{P}$ is a* probability measure*, which is a function $\mathbb{P} : \mathcal{F} \to [0,1]$ satisfying axioms $\boldsymbol{P}_1$–$\boldsymbol{P}_4$ below.*

Before formulating the axioms $\mathbf{F}_1$–$\mathbf{F}_3$ and $\mathbf{P}_1$–$\mathbf{P}_4$ we should do an example. Many of the more abstract books on probability start every section with "Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space" but we shouldn't allow ourselves to be intimidated. Here's an example to see why.

Example 1.6. *We set up a probability space to model each of the following experiments:*

1. *A single roll of a fair die in which the outcome we observe is the number thrown;*

2. *A single roll of two fair dice in which the outcome we observe is the sum of the two numbers thrown (so in particular we may not see what the individual numbers are).*

Single die. The set of outcomes of our experiment, that is our *sample space*, is $\Omega_1 = \{1, 2, 3, 4, 5, 6\}$. The *events* are all possible subsets of this; denote the set of all subsets of Ω_1 by $\mathcal{F}_1$. For example, $E_1 = \{6\}$ is the event "the result is a 6" and $E_2 = \{2, 4, 6\}$ is the event "the result is even".
We're told that the die is fair so $\mathbb{P}_1(\{i\})$ is just 1/6 and $\mathbb{P}_1(E) = \frac{1}{6}|E|$ where $|E|$ is the number of distinct elements in the subset E. Hence, $\mathbb{P}_1(E_1) = \frac{1}{6}$ and $\mathbb{P}_1(E_2) = \frac{1}{2}$.
Formally, $\mathbb{P}_1$ is a function on $\mathcal{F}_1$ which assigns a number from $[0, 1]$ to each element of $\mathcal{F}_1$.

The total on two dice. The set of outcomes that we can actually observe is $\Omega_2 = \{2, 3, 4, \ldots, 12\}$. We take $\mathcal{F}_2$ to be the set of all subsets of Ω_2. So for example $E_3 = \{2, 4, 6, 8, 10, 12\}$ is the event "the outcome is even", $E_4 = \{2, 3, 5, 7, 11\}$ is the event "the outcome is prime" and so on.
Notice now however that *not all outcomes are equally likely.* However, tabulating all possible numbers shown on the two dice we get

	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

and all of these outcomes *are* equally likely. So now we can just count to work out the probability of each event in $\mathcal{F}_2$. For example $\mathbb{P}_2(\{12\}) = \frac{1}{36}$, $\mathbb{P}_2(\{7\}) = \frac{1}{6}$, $\mathbb{P}_2(E_3) = \frac{1}{2}$ and $\mathbb{P}_2(E_4) = \frac{15}{36}$.
The probability measure is still a $[0, 1]$-valued function on $\mathcal{F}_2$, but this time it is a more interesting one.

This second example raises a very important point. The sample space that we use in modelling a particular experiment is *not unique.* In fact, to calculate the probabilities $\mathbb{P}_2$, in effect we took a larger sample space $\Omega_2' = \{(i, j) : i, j \in \{1, 2, \ldots, 6\}\}$ that records the pair of numbers thrown. But the only events that we are interested in for this particular experiment are those that tell us something about the *sum* of the numbers thrown.

In order to make sure that the theory we build is internally consistent, we need to make some assumptions about $\mathcal{F}$ and $\mathbb{P}$, in the form of axioms. To state them, we use notation from set theory.

The axioms of probability
$\mathcal{F}$ is a collection of subsets of Ω, $\mathbb{P} : \mathcal{F} \to \mathbb{R}$ and
$\mathbf{F}_1$: $\emptyset \in \mathcal{F}$, $\Omega \in \mathcal{F}$.
$\mathbf{F}_2$: If $A, B \in \mathcal{F}$, then $A^c \in \mathcal{F}$ and $A \cup B \in \mathcal{F}$.
$\mathbf{F}_3$: If $A_i \in \mathcal{F}$ for $i \geq 1$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{F}$.

$\mathbf{P}_1$: For all $A \in \mathcal{F}$, $\mathbb{P}(A) \geq 0$.
$\mathbf{P}_2$: $\mathbb{P}(\Omega) = 1$.
$\mathbf{P}_3$: If $A, B \in \mathcal{F}$ and $A \cap B = \emptyset$ then $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B)$.
$\mathbf{P}_4$: If $A_i \in \mathcal{F}$ for $i \geq 1$ and $A_i \cap A_j = \emptyset$ for $i \neq j$ then $\mathbb{P}\left(\cup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

Note that $\emptyset$ is the event that nothing happens and Ω is the event that something happens.

Axioms $\mathbf{P}_1$ and $\mathbf{P}_2$ are just to ensure that the probability of any event is a number in $[0, 1]$ and that

something happens with probability one. We already used $\mathbf{P}_3$ to calculate the probabilities of events in our examples.

In our examples, Ω was finite, so the statement about countable unions can be reduced to one about finite ones. In fact, Ω can be finite or infinite, countable or uncountable, discrete or continuous.

If Ω is a countable set, as it usually will be for the first half of this course, then we normally take $\mathcal{F}$ to be the set of all subsets of Ω (the *power set* of Ω). (You should check that, in this case, $\mathbf{F}_1$–$\mathbf{F}_3$ are satisfied.) If Ω is uncountable, however, the set of all subsets turns out to be *too large*: it ends up containing sets to which we cannot consistently assign probabilities. This is an issue which you will see discussed properly in next year's Part A Integration course; for the moment, you shouldn't worry about it, just make a mental note that there is something to be resolved here.

Axiom $\mathbf{P}_4$ will not really impinge on our consciousness. It is possible to arrange for $\mathbf{P}_1$–$\mathbf{P}_3$ to hold without $\mathbf{P}_4$, but it is hard work. Reassuringly, $\mathbf{P}_4$ follows from $\mathbf{P}_1$–$\mathbf{P}_3$ plus an intuitively appealing "continuity property":

$\mathbf{P}_4'$: If $A_1 \supseteq A_2 \supseteq \ldots$ is a sequence from $\mathcal{F}$ with $\cap_n A_n = \emptyset$, then $(\mathbb{P}(A_n))_{n\geq 1}$ is a decreasing sequence which tends to 0 as $n \to \infty$.

Example 1.7. *Consider a countable set $\Omega = \{\omega_1, \omega_2, \ldots\}$ and an arbitrary collection $(p_1, p_2, \ldots)$ of non-negative numbers with sum $\sum_{i=1}^{\infty} p_i = 1$. Put*

$$\mathbb{P}(A) = \sum_{i:\omega_i \in A} p_i.$$

Then $\mathbb{P}$ satisfies $\mathbf{P}_1$–$\mathbf{P}_4$. The numbers $(p_1, p_2, \ldots)$ are called a probability distribution.

Example 1.8. *Pick a team of m players from a squad of n, all possible teams being equally likely. Set*

$$\Omega = \left\{ (i_1, i_2, \ldots, i_n) : i_k = 0 \text{ or } 1 \text{ and } \sum_{k=1}^{n} i_k = m \right\},$$

where

$$i_k = \begin{cases} 1 & \text{if player } k \text{ is picked,} \\ 0 & \text{otherwise.} \end{cases}$$

Let $A = \{$player 1 is in the team$\}$. Then

$$\mathbb{P}(A) = \frac{\#\text{teams that include player 1}}{\#\text{possible teams}} = \frac{\binom{n-1}{m-1}}{\binom{n}{m}} = \frac{m}{n}.$$

We can derive some useful consequences of the axioms.

Theorem 1.9. *Suppose that $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and that $A, B \in \mathcal{F}$. Then*

1. $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$;
2. *If $A \subseteq B$ then $\mathbb{P}(A) \leq \mathbb{P}(B)$.*

Proof. 1. Since $A \cup A^c = \Omega$ and $A \cap A^c = \emptyset$, by $\mathbf{P}_3$, $\mathbb{P}(\Omega) = \mathbb{P}(A) + \mathbb{P}(A^c)$. By $\mathbf{P}_2$, $\mathbb{P}(\Omega) = 1$ and so $\mathbb{P}(A) + \mathbb{P}(A^c) = 1$, which entails the required result.

2. Since $A \subseteq B$, we have $B = A \cup (B \cap A^c)$. Since $B \cap A^c \subseteq A^c$, it must be disjoint from A. So by $\mathbf{P}_3$, $\mathbb{P}(B) = \mathbb{P}(A) + \mathbb{P}(B \cap A^c)$. Since by $\mathbf{P}_1$, $\mathbb{P}(B \cap A^c) \geq 0$, we thus have $\mathbb{P}(B) \geq \mathbb{P}(A)$. □

Some other useful consequences are on Problem Sheet 1.

1.4 Conditional probability

We have seen how to formalise the notion of probability. So for each event, which we thought of as an observable outcome of an experiment, we have a probability (a likelihood, if you prefer). But of course our assessment of likelihoods changes as we acquire more information and our next task is to formalise that idea. First, to get a feel for what I mean, let's look at a simple example.

Example 1.10. *Suppose that in a single roll of a fair die we know that the outcome is an even number. What is the probability that it is in fact a six?*

Solution. Let $B = \{\text{result is even}\} = \{2, 4, 6\}$ and $C = \{\text{result is a six}\} = \{6\}$. Then $\mathbb{P}(B) = \frac{1}{2}$ and $\mathbb{P}(C) = \frac{1}{6}$, but if I *know* that B has happened, then $\mathbb{P}(C|B)$ (read "the probability of C given B") is $\frac{1}{3}$ because given that B happened, we know the outcome was one of $\{2, 4, 6\}$ and since the die is fair, in the absence of any other information, we assume each of these is equally likely.

Now let $A = \{\text{result is divisible by } 3\} = \{3, 6\}$. If we know that B happened, then the only way that A can also happen is if the outcome is in $A \cap B$, in this case if the outcome is $\{6\}$ and so $\mathbb{P}(A|B) = \frac{1}{3}$ again which is $\mathbb{P}(A \cap B)/\mathbb{P}(B)$. □

Definition 1.11. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. If $A, B \in \mathcal{F}$ and $\mathbb{P}(B) > 0$ then the* conditional probability of A given B *is*

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We should check that this new notion fits with our idea of probability. The next theorem says that it does.

Theorem 1.12. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $B \in \mathcal{F}$ satisfy $\mathbb{P}(B) > 0$. Define a new function $\mathbb{Q} : \mathcal{F} \to \mathbb{R}$ by $\mathbb{Q}(A) = \mathbb{P}(A|B)$. Then $(\Omega, \mathcal{F}, \mathbb{Q})$ is also a probability space.*

Proof. Because we're using the same $\mathcal{F}$, we need only check axioms $\mathbf{P}_1$–$\mathbf{P}_4$.

$\mathbf{P}_1$. For any $A \in \mathcal{F}$,

$$\mathbb{Q}(A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \geq 0.$$

$\mathbf{P}_2$. By definition,

$$\mathbb{Q}(\Omega) = \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1.$$

$\mathbf{P}_3$ and $\mathbf{P}_4$ have the same proof, so we just do $\mathbf{P}_4$: For disjoint events $A_1, A_2, \ldots$,

$$\begin{aligned}
\mathbb{Q}(\cup_{i=1}^{\infty} A_i) &= \frac{\mathbb{P}((\cup_{i=1}^{\infty} A_i) \cap B)}{\mathbb{P}(B)} \\
&= \frac{\mathbb{P}\left(\cup_{i=1}^{\infty} (A_i \cap B)\right)}{\mathbb{P}(B)} \\
&= \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} \quad \text{(because } A_i \cap B, \, i \geq 1, \text{ are disjoint)} \\
&= \sum_{i=1}^{\infty} \mathbb{Q}(A_i).
\end{aligned}$$

□

From the definition of conditional probability, we get a very useful multiplication rule:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\,\mathbb{P}(B)\,. \tag{1.3}$$

This generalises to

$$\mathbb{P}(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2|A_1)\,\mathbb{P}(A_3|A_1 \cap A_2)\ldots\mathbb{P}(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1}) \tag{1.4}$$

(you can prove this by induction).

Example 1.13. *An urn contains 8 red balls and 4 white balls. We draw 3 balls at random* without replacement. *Let $R_i = \{$the ith ball is red$\}$ for $1 \le i \le 3$. Then*

$$\mathbb{P}(R_1 \cap R_2 \cap R_3) = \mathbb{P}(R_1)\,\mathbb{P}(R_2|R_1)\,\mathbb{P}(R_3|R_1 \cap R_2) = \frac{8}{12}\cdot\frac{7}{11}\cdot\frac{6}{10} = \frac{14}{55}.$$

Example 1.14. *A bag contains 26 tickets, one with each letter of the alphabet. If six tickets are drawn at random from the bag (without replacement), what is the chance that they can be rearranged to spell CALVIN ?*

Solution. Write A_i for the event that the ith ticket drawn is from the set $\{C, A, L, V, I, N\}$. By (1.4),

$$\mathbb{P}(A_1 \cap \ldots \cap A_6) = \frac{6}{26}\cdot\frac{5}{25}\cdot\frac{4}{24}\cdot\frac{3}{23}\cdot\frac{2}{22}\cdot\frac{1}{21}.$$ □

Example 1.15. *A bitstream when transmitted has*

$$\mathbb{P}(0 \textit{ sent}) = \frac{4}{7}, \qquad \mathbb{P}(1 \textit{ sent}) = \frac{3}{7}.$$

Owing to noise,

$$\mathbb{P}(1 \textit{ received} \mid 0 \textit{ sent}) = \frac{1}{8},$$
$$\mathbb{P}(0 \textit{ received} \mid 1 \textit{ sent}) = \frac{1}{6}.$$

What is $\mathbb{P}(0 \textit{ sent} \mid 0 \textit{ received})$*?*

Solution. Using the definition of conditional probability,

$$\mathbb{P}(0 \text{ sent} \mid 0 \text{ received}) = \frac{\mathbb{P}(0 \text{ sent and } 0 \text{ received})}{\mathbb{P}(0 \text{ received})}.$$

Now

$$\mathbb{P}(0 \text{ received}) = \mathbb{P}(0 \text{ sent and } 0 \text{ received}) + \mathbb{P}(1 \text{ sent and } 0 \text{ received}).$$

Now we use (1.3) to get

$$\begin{aligned}\mathbb{P}(0 \text{ sent and } 0 \text{ received}) &= \mathbb{P}(0 \text{ received} \mid 0 \text{ sent})\mathbb{P}(0 \text{ sent})\\ &= \big(1 - \mathbb{P}(1 \text{ received} \mid 0 \text{ sent})\big)\mathbb{P}(0 \text{ sent})\\ &= \left(1 - \frac{1}{8}\right)\frac{4}{7} = \frac{1}{2}.\end{aligned}$$

Similarly,

$$\begin{aligned}\mathbb{P}(1 \text{ sent and } 0 \text{ received}) &= \mathbb{P}(0 \text{ received} \mid 1 \text{ sent})\mathbb{P}(1 \text{ sent})\\ &= \frac{1}{6}\cdot\frac{3}{7} = \frac{1}{14}.\end{aligned}$$

Putting these together gives

$$\mathbb{P}(0 \text{ received}) = \frac{1}{2} + \frac{1}{14} = \frac{8}{14}$$

and

$$\mathbb{P}(0 \text{ sent} \mid 0 \text{ received}) = \frac{\frac{1}{2}}{\frac{8}{14}} = \frac{7}{8}.$$ □

1.5 Independence

Of course, knowing that B has happened doesn't always influence the chances of A.

Definition 1.16. *1. Events A and B are* independent *if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$.*

2. More generally, a family of events $\mathcal{A} = \{A_i : i \in I\}$ is independent *if*

$$\mathbb{P}\left(\bigcap_{i \in J} A_i\right) = \prod_{i \in J} \mathbb{P}(A_i)$$

for all **finite** *subsets J of I.*

3. A family $\mathcal{A}$ of events is pairwise independent *if $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ whenever $i \neq j$.*

WARNING: PAIRWISE INDEPENDENT DOES NOT IMPLY INDEPENDENT.

See Problem Sheet 2 for an example of this. Also, note the spelling of independent! If you use independence in solving a problem then *say so.*

Suppose that A and B are independent. Then if $\mathbb{P}(B) > 0$, we have $\mathbb{P}(A|B) = \mathbb{P}(A)$, and if $\mathbb{P}(A) > 0$, we have $\mathbb{P}(B|A) = \mathbb{P}(B)$. In other words, knowledge of the occurrence of B does not influence the probability of A, and vice versa.

Example 1.17. *Suppose we have two fair dice. Let*

$$A = \{\textit{first die shows 4}\}, \quad B = \{\textit{total score is 6}\} \quad \textit{and} \quad C = \{\textit{total score is 7}\}.$$

Then

$$\mathbb{P}(A \cap B) = \mathbb{P}(\{(4,2)\}) = \frac{1}{36}$$

but

$$\mathbb{P}(A)\,\mathbb{P}(B) = \frac{1}{6} \cdot \frac{5}{36} \neq \frac{1}{36}.$$

So A and B are not independent. However, A and C are *independent (you should check this).*

Theorem 1.18. *Suppose that A and B are independent. Then*

(a) A and B^c are independent;

(b) A^c and B^c are independent.

Proof. (a) We have $A = (A \cap B) \cup (A \cap B^c)$, where $A \cap B$ and $A \cap B^c$ are disjoint, so using the independence of A and B,

$$\mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) = \mathbb{P}(A) - \mathbb{P}(A)\,\mathbb{P}(B) = \mathbb{P}(A)\,(1 - \mathbb{P}(B)) = \mathbb{P}(A)\,\mathbb{P}(B^c).$$

(b) Apply part (a) to the events B^c and A. □

1.6 The law of total probability and Bayes' theorem

Definition 1.19. *A family of events* $\{B_1, B_2, \ldots\}$ *is a* partition *of* Ω *if*

1. $\Omega = \bigcup_{i \geq 1} B_i$ *(so that at least one* B_i *must happen), and*
2. $B_i \cap B_j = \emptyset$ *whenever* $i \neq j$ *(so that no two can happen together).*

Theorem 1.20 (The law of total probability)**.** *Suppose* $\{B_1, B_2, \ldots\}$ *is a partition of* Ω *by sets from* $\mathcal{F}$*, such that* $\mathbb{P}(B_i) > 0$ *for all* $i \geq 1$*. Then for any* $A \in \mathcal{F}$*,*

$$\mathbb{P}(A) = \sum_{i \geq 1} \mathbb{P}(A|B_i)\mathbb{P}(B_i).$$

This result is sometimes also called the *partition theorem*. We used it in our bitstream example to calculate the probability that 0 was received.

Proof. We have

$$\begin{aligned}
\mathbb{P}(A) &= \mathbb{P}(A \cap (\cup_{i \geq 1} B_i)), \text{ since } \cup_{i \geq 1} B_i = \Omega \\
&= \mathbb{P}(\cup_{i \geq 1} (A \cap B_i)) \\
&= \sum_{i \geq 1} \mathbb{P}(A \cap B_i), \text{ since } A \cap B_i,\ i \geq 1 \text{ are disjoint} \\
&= \sum_{i \geq 1} \mathbb{P}(A|B_i)\,\mathbb{P}(B_i).
\end{aligned}$$

□

Example 1.21. *Crossword setter I composes* m *clues; setter II composes* n *clues. Alice's probability of solving a clue is* α *if the clue was composed by setter I and* β *if the clue was composed by setter II.*

Alice chooses a clue at random. What is the probability she solves it?

Solution. Let

$$\begin{aligned}
A &= \{\text{Alice solves the clue}\} \\
B_1 &= \{\text{clue composed by setter I}\}, \\
B_2 &= \{\text{clue composed by setter II}\}.
\end{aligned}$$

Then

$$\mathbb{P}(B_1) = \frac{m}{m+n}, \quad \mathbb{P}(B_2) = \frac{n}{m+n}, \quad \mathbb{P}(A|B_1) = \alpha, \quad \mathbb{P}(A|B_2) = \beta.$$

By the law of total probability,

$$\mathbb{P}(A) = \mathbb{P}(A|B_1)\,\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\,\mathbb{P}(B_2) = \frac{\alpha m}{m+n} + \frac{\beta n}{m+n} = \frac{\alpha m + \beta n}{m+n}.$$

□

In our solution to Example 1.15, we combined the law of total probability with the definition of conditional probability. In general, this technique has a name:

Theorem 1.22 (Bayes' Theorem)**.** *Suppose that* $\{B_1, B_2, \ldots\}$ *is a partition of* Ω *by sets from* $\mathcal{F}$ *such that* $\mathbb{P}(B_i) > 0$ *for all* $i \geq 1$*. Then for any* $A \in \mathcal{F}$ *such that* $\mathbb{P}(A) > 0$*,*

$$\mathbb{P}(B_k|A) = \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\sum_{i \geq 1} \mathbb{P}(A|B_i)\mathbb{P}(B_i)}.$$

Proof. We have

$$\begin{aligned}\mathbb{P}(B_k|A) &= \frac{\mathbb{P}(B_k \cap A)}{\mathbb{P}(A)} \\ &= \frac{\mathbb{P}(A|B_k)\mathbb{P}(B_k)}{\mathbb{P}(A)}.\end{aligned}$$

Now substitute for $\mathbb{P}(A)$ using the law of total probability. □

See Problem Sheet 2 for a typical application of Bayes' theorem.

In Example 1.15, we calculated $\mathbb{P}(0 \text{ sent} \mid 0 \text{ received})$ by taking $\{B_1, B_2, \ldots\}$ to be $B_1 = \{0 \text{ sent}\}$ and $B_2 = \{1 \text{ sent}\}$ and A to be the event $\{0 \text{ received}\}$.

Example 1.23. *Recall Alice, from Example 1.21. Suppose that she chooses a clue at random and solves it. What is the probability that the clue was composed by setter I?*

Solution. Using Bayes' theorem,

$$\begin{aligned}\mathbb{P}(B_1|A) &= \frac{\mathbb{P}(A|B_1)\mathbb{P}(B_1)}{\mathbb{P}(A|B_1)\mathbb{P}(B_1) + \mathbb{P}(A|B_2)\mathbb{P}(B_2)} \\ &= \frac{\frac{\alpha m}{m+n}}{\frac{\alpha m}{m+n} + \frac{\beta n}{m+n}} \\ &= \frac{\alpha m}{\alpha m + \beta n}.\end{aligned}$$

□

1.7 Some useful rules for calculating probabilities

If you're faced with a probability calculation you don't know how to do, here are some things to try.

- **AND**: Try using the multiplication rule:

  $$\mathbb{P}(A \cap B) = \mathbb{P}(A|B)\,\mathbb{P}(B) = \mathbb{P}(B|A)\,\mathbb{P}(A)$$

  or its generalisation:

  $$\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n) = \mathbb{P}(A_1)\,\mathbb{P}(A_2|A_1)\ldots\mathbb{P}(A_n|A_1 \cap A_2 \cap \ldots \cap A_{n-1})$$

  (as long as all of the conditional probabilities are defined).

- **OR**: If the events are disjoint, use

  $$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \mathbb{P}(A_1) + \mathbb{P}(A_2) + \cdots + \mathbb{P}(A_n).$$

  Otherwise, try taking complements:

  $$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = 1 - \mathbb{P}((A_1 \cup A_2 \cup \ldots \cup A_n)^c) = 1 - \mathbb{P}(A_1^c \cap A_2^c \cap \ldots \cap A_n^c)$$

  ("the probability at least one of the events occurs is 1 minus the probability that none of them occur"). If that's no use, try using the inclusion-exclusion formula (see Problem Sheet 1):

  $$\mathbb{P}(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \cdots + (-1)^{n+1}\mathbb{P}(A_1 \cap A_2 \cap \ldots \cap A_n).$$

- If you can't calculate the probability of your event directly, try splitting it up according to some partition of Ω and using the law of total probability.

> ALWAYS CHECK THAT THE PROBABILITY THAT YOU CALCULATE IS IN THE INTERVAL $[0, 1]$!

Chapter 2

Discrete random variables

Interesting information about the outcome of an experiment can often be encoded as a number. For example, suppose that I am modelling the arrival of telephone calls at an exchange. Modelling this directly could be very complicated: my sample space should include all of the possible starting and finishing times of calls, all possible numbers of calls and so on. But if I am just interested in the number of calls that arrive in some time interval $[0, t]$, then I can take my sample space to be just $\Omega = \{0, 1, 2, \ldots\}$. We'll return to this example later.

Even if we are not counting something, we may be able to *encode* the result of an experiment as a number. As a trivial example, the result of a flip of a coin can be coded by letting 1 denote "head" and 0 denote "tail", say.

Real-valued discrete random variables are essentially real-valued measurements of this kind. Here's a formal definition.

Definition 2.1. *A* discrete random variable X *on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a function* $X : \Omega \to \mathbb{R}$ *such that*

(a) $\{\omega \in \Omega : X(\omega) = x\} \in \mathcal{F}$ *for each* $x \in \mathbb{R}$,

(b) $\mathrm{Im}X := \{X(\omega) : \omega \in \Omega\}$ *is a finite or countable subset of* $\mathbb{R}$.

We often abbreviate "random variable" to "r.v.".

This looks very abstract, so give yourself a moment to try to understand what it means.

- (a) says that $\{w \in \Omega : X(\omega) = x\}$ is an event to which we can assign a probability. We will usually abbreviate this event to $\{X = x\}$ and write $\mathbb{P}(X = x)$ to mean $\mathbb{P}(\{\omega \in \Omega : X(\omega) = x\})$. If these abbreviations confuse you at first, put in the ω's to make it clearer what is meant.
- (b) says that X can only take countably many values. Often $\mathrm{Im}X$ will be some subset of $\mathbb{N}$.
- If Ω is countable, (b) holds automatically because we can think of $\mathrm{Im}X$ as being indexed by Ω, and so, therefore, $\mathrm{Im}X$ must itself be countable. If we also take $\mathcal{F}$ to be the set of all subsets of Ω then (a) is also immediate.

- Later in the course, we will deal with continuous random variables, which take uncountably many values; we have to be a a bit more careful about what the correct analogue of (a) is; we will end up requiring that sets of the form $\{X \leq x\}$ are events to which we can assign probabilities.

Example 2.2. *Roll two dice and take* $\Omega = \{(i,j) : 1 \leq i,j \leq 6\}$. *Take*

$$X(i,j) = \max\{i,j\}, \quad \textit{the maximum of the two scores}$$
$$Y(i,j) = i + j, \quad \textit{the total score.}$$

A given probability space has lots of random variables associated with it. So, for example, in our telephone exchange we might have taken the "time in minutes until the arrival of the third call" in place of the number of calls by time t, say.

Definition 2.3. *The* probability mass function *(p.m.f.) of* X *is the function* $p_X : \mathbb{R} \to [0,1]$ *defined by*

$$p_X(x) = \mathbb{P}(X = x).$$

If $x \notin \mathrm{Im}X$ (that is, $X(\omega)$ never equals x) then $p_X(x) = \mathbb{P}(\{\omega : X(\omega) = x\}) = \mathbb{P}(\emptyset) = 0$. Also

$$\begin{aligned}\sum_{x \in \mathrm{Im}X} p_X(x) &= \sum_{x \in \mathrm{Im}X} \mathbb{P}(\{\omega : X(\omega) = x\}) \\ &= \mathbb{P}\left(\bigcup_{x \in \mathrm{Im}X} \{\omega : X(\omega) = x\}\right) \text{ since the events are disjoint} \\ &= \mathbb{P}(\Omega) \text{ since every } \omega \in \Omega \text{ gets mapped somewhere in } \mathrm{Im}X \\ &= 1.\end{aligned}$$

Example 2.4. *Fix an event* $A \in \mathcal{F}$ *and let* $X : \Omega \to \mathbb{R}$ *be the function given by*

$$X(\omega) = \begin{cases} 1 & \textit{if } \omega \in A, \\ 0 & \textit{otherwise.} \end{cases}$$

Then X *is a random variable with probability mass function*

$$p_X(0) = \mathbb{P}(X = 0) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A), \qquad p_X(1) = \mathbb{P}(X = 1) = \mathbb{P}(A)$$

and $p_X(x) = 0$ *for all* $x \neq 0, 1$. *We will usually write* $X = \mathbb{1}_A$ *and call this the* indicator function *of the event* A.

Notice that given a probability mass function p_X, we can always write down a probability space and a random variable defined on it with that probability mass function. For simplicity, suppose that $\mathrm{Im}X = \{0, 1, \ldots\}$. Then let $\Omega = \{0, 1, \ldots\}$, let $\mathcal{F}$ be the power set of Ω, set

$$\mathbb{P}(\{\omega\}) = p_X(\omega) \quad \text{for each } \omega \in \Omega$$

and then take X to be the identity function i.e. $X(\omega) = \omega$. However, this is often not the most natural probability space to take. For example, suppose that X represents the number of heads obtained in a sequence of three fair coin tosses. Then we could proceed as just outlined. But we could also take $\Omega = \{(i,j,k) : i,j,k \in \{0,1\}\}$, with a 0 representing a tail and a 1 representing a head, so that an element of Ω tells us exactly what the three coin tosses were. Then take $\mathcal{F}$ to be the power set of Ω,

$$\mathbb{P}(\{(i,j,k)\}) = 2^{-3} \quad \text{for all } i,j,k \in \{0,1\},$$

so that every sequence of coin tosses is equally likely, and finally set $X((i,j,k)) = i + j + k$. In both cases, X has the same distribution, but the probability spaces are quite different.

Although in our examples so far, the sample space has been explicitly present, we *can* and *will* talk about random variables X without mentioning Ω.

2.1 Some classical distributions

Before introducing concepts related to discrete random variables, we introduce a stock of examples to try these concepts out on. All are classical and ubiquitous in probabilistic modelling. They also have beautiful mathematical structure, some of which we'll uncover over the course of the term.

1. **The Bernoulli distribution.** X has the Bernoulli distribution with parameter p (where $0 \leq p \leq 1$) if
$$\mathbb{P}(X = 0) = 1 - p, \quad \mathbb{P}(X = 1) = p.$$
We often write $q = 1 - p$. (Of course since $(1 - p) + p = 1$, we must have $\mathbb{P}(X = x) = 0$ for all other values of x.) We write $X \sim \text{Ber}(p)$.

   We showed in Example 2.4 that the indicator function $\mathbb{1}_A$ of an event A is an example of a Bernoulli random variable with parameter $p = \mathbb{P}(A)$, constructed on an explicit probability space.

   The Bernoulli distribution is used to model, for example, the outcome of the flip of a coin with "1" representing heads and "0" representing tails. It is also a basic building block for other classical distributions.

2. **The binomial distribution.** X has a binomial distribution with parameters n and p (where n is a positive integer and $p \in [0, 1]$) if
$$\mathbb{P}(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}.$$
   We write $X \sim \text{Bin}(n, p)$.

   X models the number of heads obtained in n independent coin flips, where p is the probability of a head. To see this, note that the probability of any particular sequence of length n of heads and tails containing exactly k heads is $p^k(1 - p)^{n-k}$ and there are exactly $\binom{n}{k}$ such sequences.

3. **The geometric distribution.** X has a geometric distribution with parameter p
$$\mathbb{P}(X = k) = p(1 - p)^{k-1}, \quad k = 1, 2, \ldots.$$
   Notice that now X takes values in a countably infinite set – the whole of the positive integers. We write $X \sim \text{Geom}(p)$.

   We can use X to model the number of independent trials needed until we see the first success, where p is the probability of success on a single trial.

   WARNING: there is an alternative and also common definition for the geometric distribution as the distribution of the number of failures, Y, before the first success. This corresponds to $X - 1$ and so
$$\mathbb{P}(Y = k) = p(1 - p)^k, \quad k = 0, 1, \ldots.$$
   If in doubt, state which one you are using.

4. **The Poisson distribution.** X has the Poisson distribution with parameter $\lambda \geq 0$ if
$$\mathbb{P}(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k = 0, 1, \ldots.$$
   We write $X \sim \text{Po}(\lambda)$.

   This distribution arises in many applications. For example, the number of calls to arrive at a telephone exchange in a given time period or the number of electrons emitted by a radioactive source in a given time and so on. It can be extended, as we'll see, to something that evolves with time. The other setting in which we encounter it is as an approximation to a binomial distribution with a large number of trials but a low success probability for each one (see Problem Sheet 2).

Exercise 2.5. *Check that each of these really does define a probability mass function. That is:*

- $p_X(x) \geq 0$ *for all* x,
- $\sum_x p_X(x) = 1$.

Given any function p_X which is non-zero for only a finite or countably infinite number of values x and satisfying these two conditions we can define the corresponding discrete random variable – we have not produced an exhaustive list!

2.2 Expectation

By plotting the probability mass funtion for the different random variables, we get some idea of how each one will behave, but often such information can be difficult to parse and we'd like what a statistician would call "summary statistics" to give us a feel for how they behave.

The first summary statistic tells us the "average outcome" of our experiment.

Definition 2.6. *The* expectation *(or* expected value *or* mean*) of* X *is*

$$\mathbb{E}[X] = \sum_{x \in \mathrm{Im} X} x\mathbb{P}(X = x) \tag{2.1}$$

provided that $\sum_{x\in\mathrm{Im}X} |x|\mathbb{P}(X = x) < \infty$. *If* $\sum_{x\in\mathrm{Im}X} |x|\mathbb{P}(X = x)$ *diverges, we say that the* expectation does not exist.

The reason we insist that $\sum_{x\in\mathrm{Im}X} |x|\mathbb{P}(X = x)$ is finite, that is that the sum on the right-hand side of equation (2.1) is *absolutely convergent*, is that we need the expectation to take the same value regardless of the order in which we sum the terms.

Exercise 2.7. *Write down the probability mass function of a random variable whose expectation does* not *exist.*

The expectation of X is the 'average' value which X takes – if we repeat the experiment that X describes many times and take the average of the outcomes then we should expect that average to be close to $\mathbb{E}[X]$.

Example 2.8. *1. Suppose that* X *is the number obtained when we roll a fair die. Then*

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot \mathbb{P}(X = 1) + 2 \cdot \mathbb{P}(X = 2) + \ldots + 6 \cdot \mathbb{P}(X = 6)\\ &= 1 \cdot \frac{1}{6} + 2 \cdot \frac{1}{6} + \ldots + 6 \cdot \frac{1}{6} = 3.5.\end{aligned}$$

Of course, you'll never throw 3.5 on a single roll of a die, but if you throw a lot of times you expect the average number thrown to be close to 3.5. We'll come back to this idea.

2. Suppose $A \in \mathcal{F}$ *is an event and* $\mathbb{1}_A$ *is its indicator function. Then*

$$\mathbb{E}\left[\mathbb{1}_A\right] = 0 \cdot \mathbb{P}\left(A^c\right) + 1 \cdot \mathbb{P}\left(A\right) = \mathbb{P}\left(A\right).$$

3. Suppose that $\mathbb{P}\left(X = n\right) = \frac{6}{\pi^2}\frac{1}{n^2}$, $n \geq 1$. *Then*

$$\sum_{n=1}^{\infty} n\mathbb{P}\left(X = n\right) = \frac{6}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n} = \infty$$

and so the expectation does not exist.

4. Let $X \sim \mathrm{Po}(\lambda)$. Then

$$\begin{aligned}\mathbb{E}[X] &= \sum_{k=0}^{\infty} k\frac{e^{-\lambda}\lambda^k}{k!} \\ &= e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k-1)!} \\ &= \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} \\ &= \lambda e^{-\lambda}e^{\lambda} \\ &= \lambda.\end{aligned}$$

Exercise 2.9. *(On Problem Sheet 3.) Calculate the mean of the binomial and geometric distributions.*

The definition of expectation is more general than you might initially think. Let $f : \mathbb{R} \to \mathbb{R}$. Then if X is a discrete random variable, $Y = f(X)$ is also a discrete random variable.

Theorem 2.10. *If $f : \mathbb{R} \to \mathbb{R}$, then*

$$\mathbb{E}[f(X)] = \sum_{x\in \mathrm{Im}X} f(x)\mathbb{P}(X = x)$$

provided that $\sum_{x\in\mathrm{Im}X} |f(x)|\mathbb{P}(X = x) < \infty$.

Proof. Let $A = \{y : y = f(x) \text{ for some } x \in \mathrm{Im}X\}$. Then, starting from the right-hand side,

$$\begin{aligned}\sum_{x\in\mathrm{Im}X} f(x)\mathbb{P}(X = x) &= \sum_{y\in A}\sum_{x\in\mathrm{Im}X:f(x)=y} f(x)\mathbb{P}(X = x) \\ &= \sum_{y\in A}\sum_{x\in\mathrm{Im}X:f(x)=y} y\mathbb{P}(X = x) \\ &= \sum_{y\in A} y\sum_{x\in\mathrm{Im}X:f(x)=y}\mathbb{P}(X = x) \\ &= \sum_{y\in A} y\mathbb{P}(f(X) = y) \\ &= \mathbb{E}[f(X)].\end{aligned}$$

□

Example 2.11. *Take $f(x) = x^k$. Then $\mathbb{E}[X^k]$ is called the k*th moment *of X, when it exists.*

Let us now prove some properties of the expectation which will be useful to us later on.

Theorem 2.12. *Let X be a discrete random variable such that $\mathbb{E}[X]$ exists.*

(a) If X is non-negative then $\mathbb{E}[X] \geq 0$.

(b) If $a, b \in \mathbb{R}$ then $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$.

Proof. (a) We have $\mathrm{Im}X \subseteq [0, \infty)$ and so

$$\mathbb{E}[X] = \sum_{x\in\mathrm{Im}X} x\mathbb{P}(X = x)$$

is a sum whose terms are all non-negative and so must itself be non-negative.

(b) Exercise. □

The problem with using the expectation as a summary statistic is that it is too blunt an instrument in many circumstances. For example, suppose that you are investing in the stock market. If two different stocks increase at about the same rate on the average, you may still not consider them to be equally good investments. You'd like to also know something about the size of the fluctuations about that average rate.

Definition 2.13. *For a discrete random variable X, the* variance *of X is defined by*

$$\mathrm{var}\,(X) = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

provided that this quantity exists.

(This is $\mathbb{E}\,[f(X)]$ where f is given by $f(x) = (x - \mathbb{E}\,[X])^2$ – remember that $\mathbb{E}\,[X]$ is just a number.)

Note that, since $(X - \mathbb{E}\,[X])^2$ is a non-negative random variable, by part (a) of Theorem 2.12, $\mathrm{var}\,(X) \geq 0$. The variance is a measure of how much the distribution of X is spread out about its mean: the more the distribution is spread out, the larger the variance. If X is, in fact, deterministic (i.e. $\mathbb{P}\,(X = a) = 1$ for some $a \in \mathbb{R}$) then $\mathbb{E}\,[X] = a$ also and so $\mathrm{var}\,(X) = 0$: *only randomness gives rise to variance.*

Writing $\mu = \mathbb{E}[X]$ and expanding the square we see that

$$\begin{aligned}
\mathrm{var}\,(X) &= \mathbb{E}\left[(X-\mu)^2\right] \\
&= \sum_{x \in \mathrm{Im}X} (x^2 - 2\mu x + \mu^2) p_X(x) \\
&= \sum_{x \in \mathrm{Im}X} x^2 p_X(x) + 2\mu \sum_{x \in \mathrm{Im}X} x p_X(x) + \mu^2 \sum_{x \in \mathrm{Im}X} p_X(x) \\
&= \mathbb{E}\left[X^2\right] - 2\mu\mathbb{E}\,[X] + \mu^2 \\
&= \mathbb{E}\left[X^2\right] - (\mathbb{E}\,[X])^2.
\end{aligned}$$

This is often an easier expression to work with.

Those of you who have done statistics at school will have seen the *standard deviation*, which is $\sqrt{\mathrm{var}\,(X)}$. In probability, we usually work with the variance instead because it has natural mathematical properties.

Theorem 2.14. *Suppose that X is a discrete random variable whose variance exists. Then if a and b are (finite) fixed real numbers, then the variance of the discrete random variable $Y = aX + b$ is given by*

$$\mathrm{var}\,(Y) = \mathrm{var}\,(aX + b) = a^2 \mathrm{var}\,(X)\,.$$

The proof is an exercise, but notice that of course b doesn't come into it because it simply shifts the whole distribution – and hence the mean – by b, whereas variance measures relative to the mean. In view of this, why do you think statisticians often prefer to use the standard deviation rather than variance as a measure of spread?

2.3 Conditional distributions

Back in §1.4 we talked about conditional probability $\mathbb{P}(A|B)$. In the same way, for a discrete random variable X we can define its conditional distribution. This is the obvious thing: the mass function

obtained by conditioning on the outcome B.

Definition 2.15. *Suppose that B is an event such that $\mathbb{P}(B) > 0$. Then the* conditional distribution of X given B *is*

$$\mathbb{P}(X = x|B) = \frac{\mathbb{P}(\{X = x\} \cap B)}{\mathbb{P}(B)},$$

for $x \in \mathbb{R}$. The conditional expectation of X given B *is*

$$\mathbb{E}[X|B] = \sum_x x\mathbb{P}(X = x|B),$$

whenever the sum converges absolutely. We write $p_{X|B}(x) = \mathbb{P}(X = x|B)$.

Theorem 2.16 (Law of total probability for expectations)**.** *If $\{B_1, B_2, \ldots\}$ is a partition of Ω such that $\mathbb{P}(B_i) > 0$ for all $i \geq 1$ then*

$$\mathbb{E}[X] = \sum_{i \geq 1} \mathbb{E}[X|B_i]\,\mathbb{P}(B_i),$$

whenever $\mathbb{E}[X]$ exists.

Proof.

$$\begin{aligned}
\mathbb{E}[X] &= \sum_x x\mathbb{P}(X = x) \\
&= \sum_x x \left(\sum_i \mathbb{P}(X = x|B_i)\mathbb{P}(B_i) \right) \text{ by the law of total probability} \\
&= \sum_x \sum_i x\mathbb{P}(X = x|B_i)\mathbb{P}(B_i) \\
&= \sum_i \mathbb{P}(B_i) \left(\sum_x x\mathbb{P}(X = x|B_i) \right) \\
&= \sum_i \mathbb{E}[X|B_i]\mathbb{P}(B_i).
\end{aligned}$$

□

Example 2.17. *Let X be the number of rolls of a fair die required to get the first 6. (So X is geometrically distributed with parameter $1/6$.) Find $\mathbb{E}[X]$ and $\mathrm{var}(X)$.*

Solution. Let B_1 be the event that the first roll of the die gives a 6, so that B_1^c is the event that it does not. Then

$$\begin{aligned}
\mathbb{E}[X] &= \mathbb{E}[X|B_1]\mathbb{P}(B_1) + \mathbb{E}[X|B_1^c]\mathbb{P}(B_1^c) \\
&= \frac{1}{6} + \frac{5}{6}\mathbb{E}[1 + X] \quad \text{(successive rolls are independent)} \\
&= \frac{1}{6} + \frac{5}{6}(1 + \mathbb{E}[X]).
\end{aligned}$$

Rearrange to get $\mathbb{E}[X] = 6$ (as our intuition would have us guess). Similarly,

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}[X^2|B_1]\mathbb{P}(B_1) + \mathbb{E}[X^2|B_1^c]\mathbb{P}(B_1^c) \\
&= \frac{1}{6} + \frac{5}{6}\mathbb{E}[(1 + X)^2] \\
&= \frac{1}{6} + \frac{5}{6}(1 + 2\mathbb{E}[X] + \mathbb{E}[X^2]).
\end{aligned}$$

Rearranging and using the previous result ($\mathbb{E}[X] = 6$) gives $\mathbb{E}[X^2] = 66$ and so $\text{var}\,(X) = 30$.

Compare this solution to a direct calculation using the probability mass function:

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} kpq^{k-1}, \quad \mathbb{E}[X^2] = \sum_{k=1}^{\infty} k^2 pq^{k-1},$$

with $p = \frac{1}{6}$ and $q = \frac{5}{6}$. □

We'll see a powerful approach to moment calculations in §5, but first we must find a way to deal with more than one random variable at a time.

2.4 Joint distributions

Suppose that we want to consider two discrete random variables, X and Y, defined on the same probability space. In the same way as a single random variable was characterised in terms of its probability mass function, $p_X(x)$ for $x \in \mathbb{R}$, so now we must specify $p_{X,Y}(x, y) = \mathbb{P}(X = x, Y = y)$. It's not enough to specify $\mathbb{P}\,(X = x)$ and $\mathbb{P}\,(Y = y)$ because the events $\{X = x\}$ and $\{Y = y\}$ might not be independent (think of the case $Y = X^2$, for example).

Definition 2.18. *Given two random variables X and Y their* joint distribution *(or* joint probability mass function*) is*

$$p_{X,Y}(x, y) = \mathbb{P}\left(\{X = x\} \cap \{Y = y\}\right), \qquad x, y \in \mathbb{R}.$$

We usually write the right-hand side simply as $\mathbb{P}(X = x, Y = y)$. We have $p_{X,Y}(x, y) \geq 0$ for all $x, y \in \mathbb{R}$ and $\sum_x \sum_y p_{X,Y}(x, y) = 1$. The marginal distribution *of X is*

$$p_X(x) = \sum_y p_{X,Y}(x, y)$$

and the marginal *distribution of Y is*

$$p_Y(y) = \sum_x p_{X,Y}(x, y).$$

The marginal distribution of X tells you what the distribution of X is if you have no knowledge of Y.

We can write the joint mass function as a table.

Example 2.19. *Suppose that X and Y take only the values 0 or 1 and their joint mass function is given by*

Y \ X	*0*	*1*
0	$\frac{1}{3}$	$\frac{1}{2}$
1	$\frac{1}{12}$	$\frac{1}{12}$

Observe that $\sum_{x,y} p_{X,Y}(x, y) = 1$ (always a good check when modelling).

The marginals are found by summing the rows and columns:

Y \ X	*0*	*1*	$p_Y(y)$
0	$\frac{1}{3}$	$\frac{1}{2}$	$\frac{5}{6}$
1	$\frac{1}{12}$	$\frac{1}{12}$	$\frac{1}{6}$
$p_X(x)$	$\frac{5}{12}$	$\frac{7}{12}$	

Notice that $\mathbb{P}(X=1)=\frac{7}{12}$, $\mathbb{P}(Y=1)=\frac{1}{6}$ *and* $\mathbb{P}(X=1,Y=1)=\frac{1}{12}\neq\frac{7}{12}\times\frac{1}{6}$ *so* $\{X=1\}$ *and* $\{Y=1\}$ *are* not *independent events.*

Whenever $p_X(x)>0$ for some $x\in\mathbb{R}$, we can also write down the *conditional distribution of* Y *given that* $X=x$:

$$p_{Y|X=x}(y)=\mathbb{P}\left(Y=y|X=x\right)=\frac{p_{X,Y}(x,y)}{p_X(x)}\quad\text{for } y\in\mathbb{R}.$$

The *conditional expectation of* Y *given that* $X=x$ is then

$$\mathbb{E}\left[Y|X=x\right]=\sum_y y p_{Y|X=x}(y),$$

whenever the sum converges absolutely.

Example 2.20. *For the joint distribution in Example 2.19, we have*

$$p_{Y|X=0}(0)=\frac{4}{5},\quad p_{Y|X=0}(1)=\frac{1}{5}$$

and

$$\mathbb{E}\left[Y|X=0\right]=\frac{1}{5}.$$

Definition 2.21. *Discrete random variables* X *and* Y *are* independent *if*

$$\mathbb{P}(X=x,Y=y)=\mathbb{P}(X=x)\mathbb{P}(Y=y)\quad\textit{for all } x,y\in\mathbb{R}.$$

In other words, X and Y are independent if and only of the events $\{X=x\}$ and $\{Y=y\}$ are independent **for all** choices of x and y. We can also write this as

$$p_{X,Y}(x,y)=p_X(x)p_Y(y)\quad\text{for all } x,y\in\mathbb{R}.$$

Example 2.22 (Part of an old Mods question)**.** *A coin when flipped shows heads with probability* p *and tails with probability* $q=1-p$*. It is flipped repeatedly. Assume that the outcome of different flips is independent. Let* U *be the length of the initial run and* V *the length of the second run. Find* $\mathbb{P}(U=m,V=n)$, $\mathbb{P}(U=m)$, $\mathbb{P}(V=m)$*. Are* U *and* V *independent?*

Solution. We condition on the outcome of the first flip and use the law of total probability.

$$\begin{aligned}
\mathbb{P}(U=m,V=n) &= \mathbb{P}(U=m,V=n \mid \text{1st flip H})\mathbb{P}(\text{1st flip H}) + \mathbb{P}(U=m,V=n \mid \text{1st flip T})\mathbb{P}(\text{1st flip T}) \\
&= pp^{m-1}q^n p + qq^{m-1}p^n q \\
&= p^{m+1}q^n + q^{m+1}p^n. \\
\mathbb{P}(U=m) &= \sum_{n=1}^{\infty}(p^{m+1}q^n + q^{m+1}p^n) = p^{m+1}\frac{q}{1-q} + q^{m+1}\frac{p}{1-p} \\
&= p^m q + q^m p. \\
\mathbb{P}(V=n) &= \sum_{m=1}^{\infty}(p^{m+1}q^n + q^{m+1}p^n) = q^n\frac{p^2}{1-p} + p^n\frac{q^2}{1-q} \\
&= p^2q^{n-1} + q^2p^{n-1}.
\end{aligned}$$

We have $\mathbb{P}(U=m,V=n) \neq f(m)g(n)$ unless $p=q=\frac{1}{2}$. So U, V are not independent unless $p=\frac{1}{2}$. To see why, suppose that $p<\frac{1}{2}$, then knowing that U is small, say, tells you that the first run is more likely to be a run of H's and so V is likely to be longer. Conversely, knowing that U is big will tell us that V is likely to be small. U and V are *negatively correlated*. □

In the same way as we defined expectation for a single discrete random variable, so in the bivariate case we can define expectation of any function of the random variables X and Y. Let $f:\mathbb{R}^2\to\mathbb{R}$. Then

$$\begin{aligned}
\mathbb{E}[f(X,Y)] &= \sum_x\sum_y f(x,y)\mathbb{P}(X=x,Y=y) \\
&= \sum_x\sum_y f(x,y)p_{X,Y}(x,y),
\end{aligned}$$

provided the sum converges absolutely.

Theorem 2.23. *Suppose X and Y are discrete random variables and $a,b\in\mathbb{R}$ are constants. Then*

$$\mathbb{E}[aX+bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

provided that both $\mathbb{E}[X]$ and $\mathbb{E}[Y]$ exist.

Proof. Setting $f(x,y)=ax+by$, we have

$$\begin{aligned}
\mathbb{E}[f(X,Y)] &= \mathbb{E}[aX+bY] \\
&= \sum_x\sum_y (ax+by)p_{X,Y}(x,y) \\
&= a\sum_x\sum_y xp_{X,Y}(x,y) + b\sum_x\sum_y yp_{X,Y}(x,y) \\
&= a\sum_x x\left(\sum_y p_{X,Y}(x,y)\right) + b\sum_y y\left(\sum_x p_{X,Y}(x,y)\right) \\
&= a\sum_x xp_X(x) + b\sum_y yp_Y(y) \\
&= a\mathbb{E}[X] + b\mathbb{E}[Y].
\end{aligned}$$

□

Note: X and Y do *not* have to be independent for this to hold.

Theorem 2.24. *If X and Y are independent discrete random variables whose expectations exist, then*

$$\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y].$$

Proof. We have

$$\begin{aligned}
\mathbb{E}[XY] &= \sum_x \sum_y xy\mathbb{P}(X = x, Y = y) \\
&= \sum_x \sum_y xy\mathbb{P}(X = x)\mathbb{P}(Y = y) \quad \text{(by independence)} \\
&= \left(\sum_x x\mathbb{P}(X = x)\right)\left(\sum_y y\mathbb{P}(Y = y)\right) \\
&= \mathbb{E}[X]\mathbb{E}[Y].
\end{aligned}$$

□

Exercise 2.25. *Show that* $\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y)$ *when X and Y are independent.*

What happens when X and Y are *not* independent? It's useful to define the *covariance*,

$$\operatorname{cov}(X, Y) = \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right].$$

Notice that $\operatorname{cov}(X, X) = \operatorname{var}(X)$.

Exercise 2.26. *Check that* $\operatorname{cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$ *and that*

$$\operatorname{var}(X + Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X, Y).$$

Notice that this means that if X and Y are independent, their covariance is 0. In general, the covariance can be either positive or negative valued.

WARNING: $\operatorname{cov}(X, Y) = 0$ does not imply that X and Y are independent. See Problem Sheet 3 for an example.

Definition 2.27. *We can define* multivariate distributions *analogously:*

$$p_{X_1, X_2, \ldots, X_n}(x_1, x_2, \ldots, x_n) = \mathbb{P}(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n),$$

for $x_1, x_2, \ldots, x_n \in \mathbb{R}$, and so on.

Finally, by analogy with the way we defined independence for a sequence of events, we can define independence for a family of random variables.

Definition 2.28. *A family $\{X_i : i \in I\}$ of random variables are independent if for all* finite *sets $J \subseteq I$ and all collections $\{x_i : i \in J\}$ of real numbers,*

$$\mathbb{P}\left(\bigcap_{i \in J} \{X_i = x_i\}\right) = \prod_{i \in J} \mathbb{P}(X_i = x_i).$$

Chapter 3

Difference equations

Our next topic is not probability theory, but rather a tool that you need both to answer some probability questions in the next chapter, as well as in all sorts of other areas of mathematics. Here is a famous probability problem by way of motivation.

Example 3.1 (Gambler's ruin)**.** *A gambler repeatedly plays a game in which he wins £1 with probability p and loses £1 with probability $1-p$ (independently at each play). He must leave the casino if he loses all his money or if his fortune reaches £N (the house limit).*

What is the probability that he leaves with nothing if his initial fortune is £k?

Call the probability u_k and condition on the outcome of the first play to see that

$$u_k = \mathbb{P}(\text{ruin} \mid \text{initial fortune £}k \text{ and win 1st game})\mathbb{P}(\text{win 1st game}) + \mathbb{P}(\text{ruin} \mid \text{initial fortune £}k \text{ and lose 1st game})\mathbb{P}(\text{lose 1st game}).$$

So using *independence* of different plays this reads

$$u_k = pu_{k+1} + (1-p)u_{k-1}, \qquad 1 \le k \le N-1,$$

and, of course, $u_0 = 1$, $u_N = 0$.

This is an example of a second order difference equation; it is equations of this sort that we shall now learn how to solve.

Definition 3.2. *A* kth order linear difference equation *(or* recurrence relation*) has the form*

$$\sum_{j=0}^{k} a_j u_{n+j} = f(n) \tag{3.1}$$

with $a_0 \neq 0$ and $a_k \neq 0$, where $a_0, \ldots, a_k$ are constants independent of n. A solution *to such a difference equation is a sequence $(u_n)_{n\ge 0}$ satisfying (3.1) for all $n \ge 0$.*

You should keep in mind what you know about solving linear ordinary differential equations like

$$a\frac{d^2y}{dx^2} + b\frac{dy}{dx} + cy = f(x)$$

for the function y, since what we do here will be completely analogous.

The next theorem says that we can split the problem of finding a solution to our difference equations into two parts.

Theorem 3.3. *The general solution $(u_n)_{n\geq 0}$ (i.e. if the boundary conditions are not specified) of*

$$\sum_{j=0}^{k} a_j u_{n+j} = f(n)$$

can be written as $u_n = v_n + w_n$ where $(v_n)_{n\geq 0}$ is a particular solution to the equation and $(w_n)_{n\geq 0}$ solves the homogeneous equation

$$\sum_{j=0}^{k} a_j w_{n+j} = 0.$$

Proof. Suppose (u_n) has the suggested form and $(\tilde{u}_n)$ is another solution which may not necessarily be expressed in this form. Then

$$\sum_{j=0}^{k} a_j (u_{n+j} - \tilde{u}_{n+j}) = 0.$$

So (u_n) and $(\tilde{u}_n)$ differ by a solution (x_n) to the homogeneous equation. In particular,

$$\tilde{u}_n = v_n + (w_n + x_n),$$

which *is* of the suggested form since $(w_n + x_n)$ is clearly a solution to the homogeneous equation. □

3.1 First order linear difference equations

We will develop the necessary methods via a series of worked examples.

Example 3.4. *Solve*

$$u_{n+1} = au_n + b$$

where $u_0 = 3$ and the constants $a \neq 0$ and b are given

Solution. The homogeneous equation is $w_{n+1} = aw_n$. "Putting it into itself", we get

$$w_n = aw_{n-1} = \ldots = a^n w_0 = Aa^n$$

for some constant A.

How about a particular solution? As in differential equations, guess a constant solution might work, so try $v_n = C$. This gives $C = aC + b$ so provided that $a \neq 1$, $C = \frac{b}{1-a}$ and we have general solution

$$u_n = Aa^n + \frac{b}{1-a}.$$

Setting $n = 0$ allows us to determine A:

$$3 = A + \frac{b}{1-a} \text{ and so } A = 3 - \frac{b}{1-a}.$$

Hence,

$$u_n = \left(3 - \frac{b}{1-a}\right) a^n + \frac{b}{1-a} = 3a^n + \frac{b(1-a^n)}{1-a}.$$

What happens if $a = 1$? An applied maths-type approach would set $a = 1 + \epsilon$ and try to see what happens as $\epsilon \to 0$:

$$\begin{aligned} u_n &= u_0(1+\epsilon)^n + \frac{b(1-(1+\epsilon)^n)}{1-(1+\epsilon)} \\ &= u_0 + b\frac{(1-(1+n\epsilon))}{-\epsilon} + \mathcal{O}(\epsilon) \\ &= u_0 + nb + \mathcal{O}(\epsilon) \to u_0 + nb \quad \text{as } \epsilon \to 0. \end{aligned}$$

An alternative approach is to mimic what you did for differential equations and "try the next most complex thing". We have $u_{n+1} = u_n + b$ and the homogeneous equation has solution $w_n = A$ (a constant). For a particular solution try $v_n = Cn$ (note that there is no point in adding a constant term because the constant solves the homogeneous equation and so it makes no contribution to the right-hand side when we substitute).

Then $C(n+1) = Cn + b$ gives $C = b$ and we obtain once again the general solution

$$u_n = A + bn.$$

Setting $n = 0$ yields $A = 3$ and so $u_n = 3 + bn$. □

Example 3.5.

$$u_{n+1} = au_n + bn.$$

Solution. As above, the homogeneous equation has solution $w_n = Aa^n$. For a particular solution, try $v_n = Cn + D$. Substituting

$$C(n+1) + D = a(Cn + D) + bn.$$

Equating coefficients of n and the constant terms gives

$$C = aC + b, \quad C + D = aD,$$

so again provided $a \neq 1$ we can solve to obtain $C = \frac{b}{1-a}$ and $D = \frac{-c}{1-a}$. Thus for $a \neq 1$

$$u_n = Aa^n + \frac{bn}{1-a} - \frac{b}{(1-a)^2}.$$

To find A, we need a boundary condition (e.g. the value of u_0). □

Exercise 3.6. *Solve the equation for $a = 1$. Hint: try $v_n = Cn + Dn^2$.*

3.2 Second order linear difference equations

Consider

$$u_{n+1} + au_n + bu_{n-1} = f(n).$$

The general solution will depend on *two* constants. For the first order case, the homogeneous equation had a solution of the form $w_n = A\lambda^n$, so we try the same here. Substituting $w_n = A\lambda^n$ in

$$w_{n+1} + aw_n + bw_{n-1} = 0$$

gives

$$A\lambda^{n+1} + aA\lambda^n + bA\lambda^{n-1} = 0.$$

For a non-trivial solution we can divide by $A\lambda^{n-1}$ and see that λ must solve the quadratic equation

$$\lambda^2 + a\lambda + b = 0.$$

This is called the *auxiliary equation.* (So just as when you solve 2nd order ordinary differential equations you obtain a quadratic equation by considering solutions of the form $e^{\lambda t}$, so here we obtain a quadratic in λ by considering solutions of the form λ^n.)

If the auxiliary equation has distinct roots, λ_1 and λ_2 then the general solution to the homogeneous equation is

$$w_n = A_1\lambda_1^n + A_2\lambda_2^n.$$

If $\lambda_1 = \lambda_2 = \lambda$ try the next most complicated thing (or mimic what you do for ordinary differential equations) to get

$$w_n = (A + Bn)\lambda^n.$$

Exercise 3.7. *Check that this solution works.*

How about particular solutions? The same tricks as for the one-dimensional case apply. You can only guess, but your best guess is something of the same form as f, and if that fails try the next most complicated thing. You can save yourself work by not including components that you already know solve the homogeneous equation.

Example 3.8. *Solve*

$$u_{n+1} + 2u_n - 3u_{n-1} = 1.$$

Solution. The auxiliary equation is just

$$\lambda^2 + 2\lambda - 3 = 0$$

which has roots $\lambda_1 = -3$, $\lambda_2 = 1$, so

$$w_n = A(-3)^n + B.$$

For a particular solution, we'd like to try a constant, but that won't work because we know that the it solves the homogeneous equation (it's a special case of w_n). So try the next most complicated thing, which is $v_n = Cn$. Substituting, we obtain

$$C(n+1) + 2Cn - 3C(n-1) = 1,$$

which gives $C = \frac{1}{4}$. The general solution is then

$$u_n = A(-3)^n + B + \frac{1}{4}n.$$

If the boundary conditions had been specified, you could now find A and B by substitution. (Note that it takes one boundary condition to specify the solution to a first order difference equation and two to specify the solution to a 2nd order difference equation. Usually these will be the values of u_0 and u_1 but notice that in the gambler's ruin problem we are given u_0 and u_N.) □

One last example:

Example 3.9. *Solve*

$$u_{n+1} - 2u_n + u_{n-1} = 1.$$

Solution. The auxiliary equation $\lambda^2 - 2\lambda + 1 = 0$ has repeated root $\lambda = 1$, so the homogeneous equation has general solution

$$w_n = An + B.$$

For a particular solution, try the next most complicated thing, so $v_n = Cn^2$. (Once again there is no point in adding a $Dn + E$ term to this as that solves the homogeneous equation, so substituting it on the left cannot contribute anything to the 1 that we are trying to obtain on the right of the equation.) Substituting, we obtain

$$C(n+1)^2 - 2Cn^2 + C(n-1)^2 = 1,$$

which gives $C = \frac{1}{2}$. So the general solution is

$$u_n = An + B + \frac{1}{2}n^2.$$ □

Solution to the Gambler's ruin problem (Example 3.1). We have

$$u_k = pu_{k+1} + (1-p)u_{k-1}, \qquad 1 \le k \le N-1,$$

and $u_0 = 1$, $u_N = 0$. This is a homogeneous second-order difference equation. The auxiliary equation is

$$p\lambda^2 - \lambda + (1-p) = 0$$

which factorises as

$$(p\lambda - (1-p))(\lambda - 1) = 0.$$

So $\lambda = \frac{1-p}{p}$ or 1. If $p \neq \frac{1}{2}$ then

$$u_k = A + B\left(\frac{1-p}{p}\right)^k$$

for some constants A and B which we can find using the boundary conditions:

$$u_0 = 1 = A + B \quad \text{and} \quad u_N = 0 = A + B\left(\frac{1-p}{p}\right)^N.$$

These give

$$A = -\frac{\left(\frac{1-p}{p}\right)^N}{1 - \left(\frac{1-p}{p}\right)^N}, \quad B = \frac{1}{1 - \left(\frac{1-p}{p}\right)^N}$$

and so

$$u_k = \frac{\left(\frac{1-p}{p}\right)^k - \left(\frac{1-p}{p}\right)^N}{1 - \left(\frac{1-p}{p}\right)^N}.$$

Exercise: check that in the case $p = \frac{1}{2}$ we get

$$u_k = 1 - \frac{k}{N}, \quad 0 \le k \le N.$$ □

Chapter 4

Random walks

Imagine a particle which at each time point $n = 0, 1, 2, \ldots$ has a position in the set $\{0, 1, 2, \ldots, N\}$. For each position, there are rules determining where the particle can move to at the next time step from that position and with what probability it moves to each of the possible new positions. The important point is that these rules *only* depend on the current position, not on the earlier positions that the particle has visited. This is what we mean by a *random walk*. We will usually assume that the particle can only ever move to one of the neighbouring positions, in which case we say that the random walk is *simple*.

Random walks can be used to model various real-world situations. For example, the path traced by a molecule as it moves in a liquid or a gas; the path of an animal searching for food; or the price of a particular stock every Monday morning.

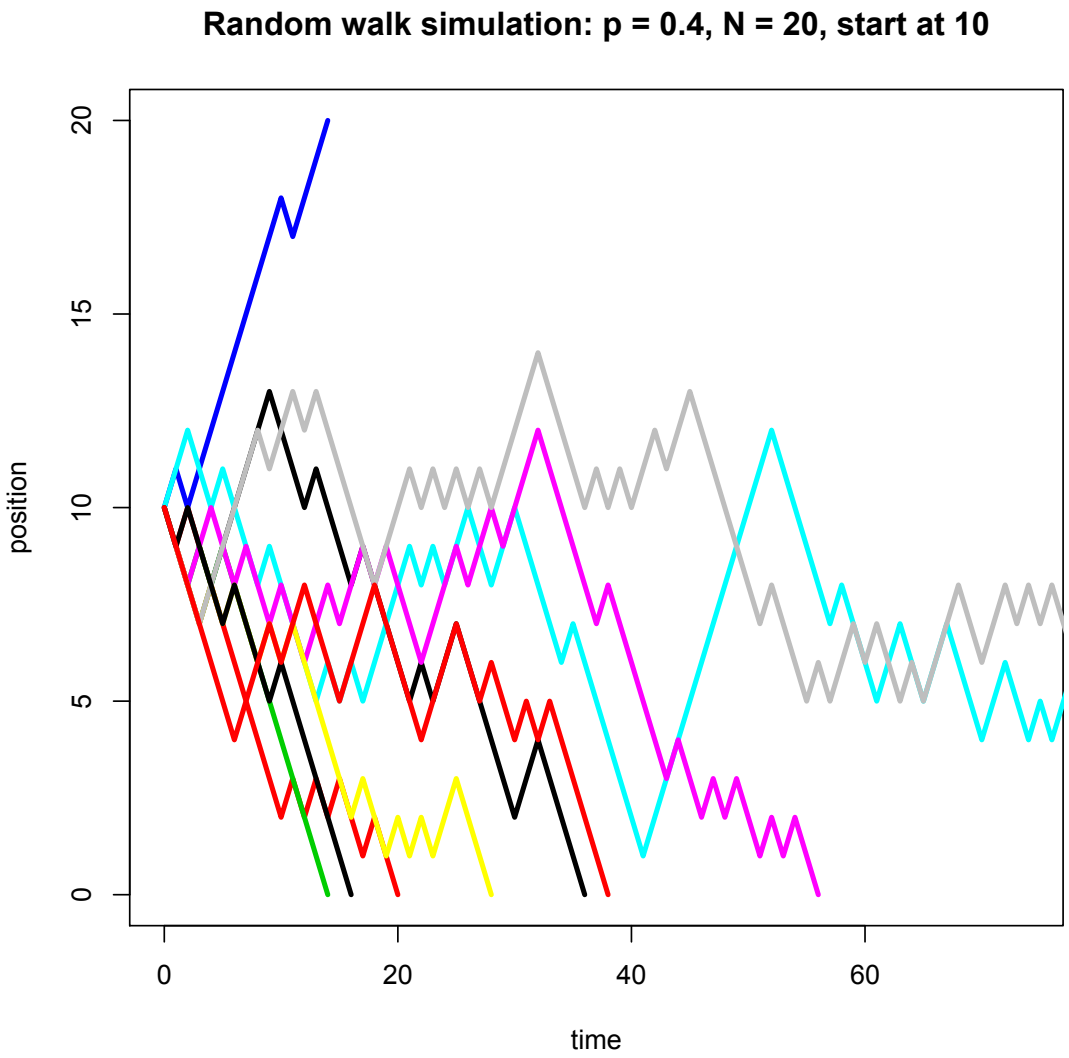
To be more definite, suppose that S_n denotes the position of the particle at time $n \geq 0$. In general, its starting position, S_0, may be random, although it will often be fixed and deterministic. We will suppose that if $S_n \in \{1, 2, \ldots, N-1\}$ for some $n \geq 0$, then

$$S_{n+1} = \begin{cases} S_n + 1 & \text{with probability } p \\ S_n - 1 & \text{with probability } q, \end{cases}$$

where $q = 1 - p$. In other words, the walk either makes an "up" step, with probability p, or a "down" step, with probability q. We have to treat the cases where the walk is at one of the end-points, i.e. $S_n = 0$ or $S_n = N$, separately. Two common rules are that the end-points 0 and N are *absorbing* (the particle stays there forever i.e. if $S_n = N$ then $S_{n+1} = N$) or *reflecting* (the particle moves to the unique neighbouring point i.e. if $S_n = N$ then $S_{n+1} = N - 1$). The simple random walk is *symmetric* if $p = q$.

Notice that we have just defined a collection $\{S_0, S_1, S_2, \ldots\}$ of random variables with certain relationships between them. Note that $S_0, S_1, \ldots$ are *not* independent: for example, if $S_5 = 4$, S_6 can only take the values 3 or 5, whereas if $S_5 = 1$, S_6 can only take the values 0 or 2. Because the index now refers to time, we prefer to call this collection of random variables a *random process* and write $(S_n)_{n \geq 0}$. The simple random walk is the only sort of random process we will deal with in this course, but you will meet many more of them in Part A Probability.

On the next page, you can see the first part of 10 simulated random walk paths on the set $\{0, 1, 2, \ldots, 20\}$ with starting point 10, absorbing barriers at 0 and 20 and "up probability" $p = 0.4$. Some paths get absorbed at 0, some get absorbed at 20 and some have not yet been absorbed at either.



For a fixed value $m \geq 0$, we can write down the joint distribution of $S_0, S_1, \ldots, S_m$ relatively easily. Suppose that the walk starts at 5 and we have absorbing barriers at 0 and 8. Then, for example,

$$\mathbb{P}(S_0 = 5, S_1 = 6, S_2 = 5) = \mathbb{P}(S_0 = 5)\,\mathbb{P}(S_1 = 6|S_0 = 5)\,\mathbb{P}(S_2 = 5|S_0 = 5, S_1 = 6)$$

(make sure you're happy with this expression – it's just an application of the multiplication rule). We said that the random walk starts at 5, so $\mathbb{P}(S_0 = 5) = 1$. Now

$$\mathbb{P}(S_1 = 6|S_0 = 5) = p.$$

Finally,

$$\mathbb{P}(S_2 = 5|S_0 = 5, S_1 = 6) = q,$$

because the step we make at time 2 is independent of all the previous steps we made. So we get

$$\mathbb{P}(S_0 = 5, S_1 = 6, S_2 = 5) = pq.$$

In the gambler's ruin problem, the wealth of the gambler performs a simple random walk with initial position $S_0 = k$ and absorbing barriers at 0 and N. We showed that the probability of absorption at 0 is

$$u_k = \begin{cases} \frac{\left(\frac{q}{p}\right)^k - \left(\frac{q}{p}\right)^N}{1-\left(\frac{q}{p}\right)^N} & \text{if } p \neq q \\ 1 - \frac{k}{N} & \text{if } p = q = 1/2, \end{cases} \qquad 0 \leq k \leq N.$$

In terms of the random walk, the event "absorption at 0" can be written as

$$\{\text{there exists } n \geq 1 \text{ such that } S_n = 0\}$$

and

$$u_k = \mathbb{P}\left(\text{there exists } n \geq 1 \text{ such that } S_n = 0 | S_0 = k\right).$$

This raises an interesting question: must the particle be absorbed at either 0 or N? Or is there some probability of it just staying in the subset $\{1, 2, \ldots, N-1\}$ forever? We can resolve this by finding

$$v_k := \mathbb{P}\left(\text{there exists } n \geq 1 \text{ such that } S_n = N | S_0 = k\right), \quad 0 \leq k \leq N.$$

The argument is completely analogous to that we used to find u_k. We obtain the same difference equation but with opposite boundary conditions, $v_0 = 0$ and $v_N = 1$. This leads to

$$v_k = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^k}{1-\left(\frac{q}{p}\right)^N} & \text{if } p \neq q \\ \frac{k}{N} & \text{if } p = q = 1/2, \end{cases} \qquad 0 \leq k \leq N.$$

and we see that

$$u_k + v_k = 1$$

for all $0 \leq k \leq N$. So the random walk is either absorbed at 0 or at N, and there are no other possibilities.

Because absorption occurs at a random time which could be arbitrarily large, it's rather difficult to find u_k and v_k directly from the joint distribution function of $S_0, S_1, \ldots$. The approach we used above, whereby we conditioned on the first step and used that to write down a difference equation, is much more effective.

We will go through a series of worked examples of random walk problems to which we can apply the probabilistic tools we have developed. In what follows, we're often going to work with probabilities conditioned on the event $\{S_0 = k\}$ for a fixed value k. Suppose, more generally, that C is an event with $\mathbb{P}(C) > 0$. Recall that the conditional probability $\mathbb{Q} : \mathcal{F} \to [0, 1]$ defined by

$$\mathbb{Q}(A) := \mathbb{P}(A|C)$$

is *itself* a probability measure. In particular, we can condition on further events:

$$\mathbb{Q}(A|B) = \frac{\mathbb{Q}(A \cap B)}{\mathbb{Q}(B)} = \frac{\mathbb{P}(A \cap B|C)}{\mathbb{P}(B|C)} = \frac{\mathbb{P}(A \cap B \cap C)}{\mathbb{P}(B \cap C)} = \mathbb{P}(A|B \cap C).$$

So, for example, we can apply the law of total probability to $\mathbb{Q}$: if $B_1, B_2, \ldots, B_n$ is a partition of Ω such that $\mathbb{Q}(B_i) > 0$ for all $1 \leq i \leq n$ then

$$\mathbb{Q}(A) = \sum_{i=1}^{n} \mathbb{Q}(A|B_i)\mathbb{Q}(B_i).$$

Converting back to $\mathbb{P}$, this says that

$$\mathbb{P}(A|C) = \sum_{i=1}^{n} \mathbb{P}(A|B_i \cap C)\,\mathbb{P}(B_i|C).$$

Similarly, if X is a discrete random variable, we get

$$\mathbb{E}[X|C] = \sum_{i=1}^{n} \mathbb{E}[X|B_i \cap C]\,\mathbb{P}(B_i|C).$$

Example 4.1. *Suppose we start at k and we know that absorption occurs at 0. What is the probability that the first move was from k to $k-1$? (Assume $p \neq q$.)*

Solution. We want to calculate $\mathbb{P}(S_1 = k-1|S_0 = k, \text{absorption at } 0)$. Manipulating the conditional probabilities, we get

$$\begin{aligned}\mathbb{P}(S_1 = k-1|S_0 = k, \text{absorption at } 0) &= \frac{\mathbb{P}(S_1 = k-1, \text{absorption at } 0|S_0 = k)}{\mathbb{P}(\text{absorption at } 0|S_0 = k)} \\ &= \frac{\mathbb{P}(S_1 = k-1|S_0 = k)\,\mathbb{P}(\text{absorption at } 0|S_0 = k, S_1 = k-1)}{\mathbb{P}(\text{absorption at } 0|S_0 = k)}.\end{aligned}$$

We already know that $\mathbb{P}(S_1 = k-1|S_0 = k) = q$ and $\mathbb{P}(\text{absorption at } 0|S_0 = k) = u_k$. What about $\mathbb{P}(\text{absorption at } 0|S_0 = k, S_1 = k-1)$? This is exactly like asking for the probability of absorption at 0 started from $k-1$ – the fact that we made a step from k to $k-1$ before that doesn't affect what happens from $k-1$ on. So this probability is equal to u_{k-1}. So we get

$$\mathbb{P}(S_1 = k-1|S_0 = k, \text{absorption at } 0) = \frac{qu_{k-1}}{u_k} = \frac{q((\frac{q}{p})^{k-1} - (\frac{q}{p})^N)}{(\frac{q}{p})^k - (\frac{q}{p})^N}.$$

Intuitively, we would expect this to be bigger than q – knowing that the walk is absorbed at 0 should make it more likely that the first step was downwards. This turns out to be true – check! (You might find it helpful to treat the cases $p > q$ and $q > p$ separately.) □

Example 4.2. *Now suppose we randomly allocate the initial point on the integers $0, 1, \ldots, N$ i.e. S_0 has a uniform distribution on $\{0, 1, \ldots, N\}$. Again assume that $p \neq q$. What is the probability that absorption occurs at 0 rather than N?*

Solution. We will apply the usual law of total probability. We take the partition to be over the starting point of the chain i.e. the partition is given by the events $\{S_0 = 0\}, \{S_0 = 1\}, \ldots, \{S_0 = N\}$. So

$$\mathbb{P}(\text{absorption at } 0) = \sum_{k=0}^{N} \mathbb{P}(\text{absorption at } 0|S_0 = k)\,\mathbb{P}(S_0 = k)$$

Uniform selection of the starting point gives $\mathbb{P}(S_0 = k) = \frac{1}{N+1}$, for $k = 0, 1, \ldots, N$. Hence,

$$\begin{aligned}\mathbb{P}(\text{absorption at } 0) &= \sum_{k=0}^{N} u_k \frac{1}{N+1} \\ &= \frac{1}{N+1} \sum_{k=0}^{N} \frac{(\frac{q}{p})^k - (\frac{q}{p})^N}{1 - (\frac{q}{p})^N} \\ &= \frac{1}{N+1} \frac{\left(\sum_{k=0}^{N} (\frac{q}{p})^k\right) - (N+1)(\frac{q}{p})^N}{1 - (\frac{q}{p})^N} \\ &= \frac{1 - (\frac{q}{p})^{N+1} - (N+1)(1 - (\frac{q}{p}))(\frac{q}{p})^N}{(N+1)(1 - (\frac{q}{p}))(1 - (\frac{q}{p})^N)} \\ &= \frac{1 - (N+1)(\frac{q}{p})^N + N(\frac{q}{p})^{N+1}}{(N+1)(1 - (\frac{q}{p}))(1 - (\frac{q}{p})^N)}.\end{aligned}$$

□

Example 4.3. *In principle, we can deal with different move possibilities and probabilities. For example, suppose that given $S_n = k \in \{1, 2, \ldots, N-2\}$ for some $n \geq 0$, the particle moves as follows:*

$$S_{n+1} = \begin{cases} S_n - 1 & \text{with probability } q, \\ S_n + 1 & \text{with probability } p, \\ S_n + 2 & \text{with probability } r, \end{cases}$$

where $q + p + r = 1$. Suppose that we have absorbing barriers at 0 and N and that if $S_n = N - 1$ then

$$S_{n+1} = \begin{cases} S_n - 1 & \text{with probability } q, \\ S_n + 1 & \text{with probability } p + r. \end{cases}$$

What is the probability of absorption at 0 given we start at k?

Solution. Using our conditional version of the law of total probability, and partitioning over the different possible moves at the first step, we get

$$\begin{aligned} \mathbb{P}(\text{absorption at } 0 | S_0 = k) = {} & \mathbb{P}(\text{absorption at } 0 | S_0 = k, S_1 = k-1)\, \mathbb{P}(S_1 = k-1 | S_0 = k) \\ & + \mathbb{P}(\text{absorption at } 0 | S_0 = k, S_1 = k+1)\, \mathbb{P}(S_1 = k+1 | S_0 = k) \\ & + \mathbb{P}(\text{absorption at } 0 | S_0 = k, S_1 = k+2)\, \mathbb{P}(S_1 = k+2 | S_0 = k) \end{aligned}$$

for $1 \leq k \leq N-2$. Let $u_k = \mathbb{P}(\text{absorption at } 0 | S_0 = k)$. Then we have

$$u_k = q u_{k-1} + p u_{k+1} + r u_{k+2}, \quad 1 \leq k \leq N-2.$$

This is a third-order difference equation. Moreover, $u_0 = 1$, $u_N = 0$ and

$$u_{N-1} = (p + r) u_N + q u_{N-2};$$

these three equations provide the necessary boundary conditions. In principle we need to solve the cubic

$$p\lambda^3 + r\lambda^2 - \lambda + q = 0$$

but since $\lambda = 1$ is clearly a root, in practice we only need to solve a quadratic. □

Example 4.4. *The number of steps until absorption (at either 0 or N) is random. Assume $p \neq q$. What is the expected number of steps to absorption?*

Solution. We will use the conditional version of the law of total probability for expectations. Let X be the number of steps (from time 0 onwards) to absorption. Then for $k \in \{1, 2, \ldots, N-1\}$,

$$\begin{aligned} \mathbb{E}[X | S_0 = k] = {} & \mathbb{E}[X | S_0 = k, S_1 = k+1]\, \mathbb{P}(S_1 = k+1 | S_0 = k) \\ & + \mathbb{E}[X | S_0 = k, S_1 = k-1]\, \mathbb{P}(S_1 = k-1 | S_0 = k). \end{aligned}$$

Set $e_k = \mathbb{E}[X | S_0 = k]$. There are two terms here we need to think carefully about: $\mathbb{E}[X | S_0 = k, S_1 = k+1]$ and $\mathbb{E}[X | S_0 = k, S_1 = k-1]$. Let $\tilde{X}$ be the number of steps from *time 1 onwards* until the walk is absorbed. Clearly, $X = 1 + \tilde{X}$, as long as $S_0 \neq 0, N$ (if $S_0 = 0$ or N, the time to absorption is just 0). Now

$$\mathbb{E}[X | S_0 = k, S_1 = k+1] = \mathbb{E}\left[1 + \tilde{X} | S_0 = k, S_1 = k+1\right] = 1 + \mathbb{E}\left[\tilde{X} | S_0 = k, S_1 = k+1\right].$$

Now, since we know that $S_1 = k+1$, the value of S_0 doesn't have any influence on $\tilde{X}$. Moreover, starting at time 1 and asking for the expected time to absorption from $k+1$ is exactly the same problem as starting at time 0 and asking for the expected time to absorption from $k+1$. So

$$\mathbb{E}\left[\tilde{X} | S_0 = k, S_1 = k+1\right] = \mathbb{E}[X | S_0 = k+1] = e_{k+1}.$$

The same argument works to show that $\mathbb{E}[X|S_0 = k, S_1 = k-1] = 1 + e_{k-1}$. So we get

$$e_k = (1 + e_{k-1})q + (1 + e_{k+1})p$$

which rearranges to give

$$pe_{k+1} - e_k + qe_{k-1} = -1,$$

with boundary conditions $e_0 = e_N = 0$.

The homogeneous equation is

$$pf_{k+1} - f_k + qf_{k-1} = 0$$

which gives auxiliary equation

$$p\lambda^2 - \lambda + q = 0$$

with solutions $\lambda = 1$ or q/p. So $f_k = A + B(q/p)^k$.

For a particular solution, try $g_k = Ck$ (it's no good trying a constant since that's part of the complementary solution). This yields

$$pC(k+1) - Ck + qC(k-1) = -1$$

and so $C = 1/(q-p)$. Putting everything together, we get

$$e_k = A + B\left(\frac{q}{p}\right)^k + \frac{k}{q-p}.$$

Using the boundary conditions, we get

$$e_0 = 0 = A + B, \quad e_N = 0 = A + B\left(\frac{q}{p}\right)^N + \frac{N}{q-p}.$$

Solving for A and B, we finally obtain

$$e_k = \frac{N}{(p-q)(1-(\frac{q}{p})^N)} - \frac{N}{(p-q)(1-(\frac{q}{p})^N)}\left(\frac{q}{p}\right)^k - \frac{k}{p-q}$$

for $0 \le k \le N$. □

Example 4.5. *Suppose we have a simple random walk with $p \neq q$, absorption at 0 and a reflecting barrier at N. What is the expected time until absorption?*

Solution. Let e_k be the expected number of steps to absorption (which can now only happen at 0). The difference equation for e_k is the same:

$$pe_{k+1} - e_k + qe_{k-1} = -1,$$

for $1 \le k \le N-1$. We still have $e_0 = 0$, but because of the reflection at N we get

$$\mathbb{E}[X|S_0 = N] = \mathbb{E}\left[1 + \tilde{X}|S_0 = N, S_1 = N-1\right]\mathbb{P}(S_1 = N-1|S_0 = N).$$

Since $\mathbb{P}(S_1 = N-1|S_0 = N) = 1$, arguing as before this gives

$$e_N = 1 + e_{N-1},$$

which is the second boundary condition. Solving as before, we get

$$e_k = A + B\left(\frac{q}{p}\right)^k + \frac{k}{q-p}$$

and the boundary conditions give

$$e_k = \frac{2p^2}{(q-p)^2}\left(\frac{p}{q}\right)^{N-1}\left(1-\left(\frac{q}{p}\right)^k\right)+\frac{k}{q-p}. \qquad \square$$

What happens if we instead have a random walk on the infinite set $\{0,1,2,\ldots\}$ with an absorbing barrier at 0? We can get an idea by letting $N\to\infty$. Let's see what happens to the probability, $u_k^{(N)}$, of absorption at 0 started from k as $N\to\infty$. We have

$$\lim_{N\to\infty} u_k^{(N)} = \begin{cases} \lim_{N\to\infty} \frac{\left(\frac{q}{p}\right)^k-\left(\frac{q}{p}\right)^N}{1-\left(\frac{q}{p}\right)^N} & \text{if } p\neq q \\ \lim_{N\to\infty} 1-\frac{k}{N} & \text{if } p=q=1/2 \end{cases} = \begin{cases} \left(\frac{q}{p}\right)^k & \text{if } p>q \\ 1 & \text{if } p\leq q. \end{cases}$$

It turns out that this really does give the absorption probability for the random walk on the infinite set. So if the probability of moving left is greater or equal to the probability of moving right at each step, the random walk *always* hits 0. On the other hand, if the probability of moving right is strictly larger than the probability of moving left then there is a strictly positive probability $(1-(q/p)^k)$ that the walk will *never* hit 0, started from k. Why is this not a rigorous argument? Because we don't know that the absorption probability for the random walk on an infinite set can be obtained by simply sending $N\to\infty$ in the finite problem – we would need some sort of continuity property of the absorption probability in order to make this argument work. The proof given below is rather involved and is therefore non-examinable, but is a nice application of some of the methods we have developed in this course so far (including solving difference equations and enumerating combinatorial objects by finding a bijection with something easier to count). There are quicker proofs of this result, but they require more mathematical technology!

Theorem 4.6. *(Non-examinable) Suppose that $(S_n)_{n\geq 0}$ a simple random walk on $\{0,1,2,\ldots\}$ with up probability p and an absorbing barrier at 0. Let*

$$u_k = \mathbb{P}\left(\textit{absorption at 0}\,|S_0=k\right), \quad k\geq 0.$$

Then

$$u_k = \begin{cases} \left(\frac{q}{p}\right)^k & \textit{if } p>q, \\ 1 & \textit{if } p\leq q. \end{cases}$$

Proof. As in the case of a random walk on $\{0,1,2,\ldots,N\}$, we get

$$u_k = pu_{k+1}+qu_{k-1}$$

which, for $p\neq q$, has general solution

$$u_k = A+B\left(\frac{q}{p}\right)^k$$

for some constants A and B and, for $p=q=1/2$, has general solution

$$u_k = C+Dk$$

for some constants C and D. We now have only a single boundary condition, $u_0=1$, which tells us that $A=1-B$ and that $C=1$. So we need an argument which will allow us to determine B and D.

Finding D is easy: u_k is a probability for all $k\geq 0$ and so, in particular, it must lie between 0 and 1. The only way to arrange for this to be true for all k is to take $D=0$ (if $D<0$ then eventually $1+Dk$

will be negative; if $D > 0$ then eventually $1 + Dk$ will exceed 1). So in the symmetric case, $u_k = 1$ for all $k \geq 0$. The same argument works to give $B = 0$ for $p < 1/2$, since $(q/p)^k$ is increasing in k.

Finding B in the case $p > q$ is a bit more involved. We will do it by calculating u_1. First notice that

$$\begin{aligned} u_1 &= \mathbb{P}\left(\bigcup_{n=0}^{\infty} \{S_{2n+1} = 0 \text{ and } S_k \geq 1, 1 \leq k \leq 2n\} \,\middle|\, S_0 = 1 \right) \\ &= \sum_{n=0}^{\infty} \mathbb{P}\left(S_{2n+1} = 0 \text{ and } S_k \geq 1, 1 \leq k \leq 2n | S_0 = 1\right). \end{aligned} \tag{$*$}$$

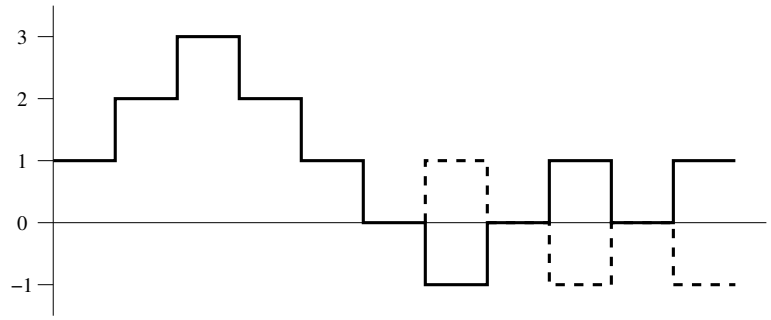
since the union is of disjoint events. We'll calculate a generic term in the sum.

Fix $n \geq 0$. Any path from $S_0 = 1$ to $S_{2n+1} = 0$ such that $S_k \geq 1$ for $1 \leq k \leq 2n$ must have $S_{2n} = 1$. Moreover, it contains n up-steps and $n+1$ down-steps and so must have probability $p^n q^{n+1}$. How many such paths are there? It's sufficient to count paths which go from 1 to 1 in $2n$ steps and which stay ≥ 1, since the last step is always down to 0. It's easier to think of this as

$$\#\{\text{paths from 1 to 1 in } 2n \text{ steps}\} - \#\{\text{paths from 1 to 1 in } 2n \text{ steps which go through 0}\}.$$

There are $\binom{2n}{n}$ paths from 1 to 1 in $2n$ steps, since we need n up-steps and n down-steps and there are $\binom{2n}{n}$ ways to choose the positions of the up-steps.

To count paths from 1 to 1 in $2n$ steps which go through 0, we use the *reflection principle*. This says that there is a bijection between these paths and paths from 1 to -1 in $2n$ steps. The easiest way to see why this is true is to draw a picture:



The solid line is a path from 1 to 1 in $2n$ steps (with $n = 6$) which goes through 0. When it first hits 0, create a new path by doing the opposite step each time to that done by the original path. This gives a path from 1 to -1 in $2n$ steps. You can check that *every* path from 1 to -1 in $2n$ steps can be obtained in this way (and so we really do have a bijection between the two sets of paths).

Fortunately, paths from 1 to -1 in $2n$ steps are easier to enumerate! There are $\binom{2n}{n+1}$ of them, since we just need to choose the positions of the $n+1$ down-steps. So we obtain

$$\#\{\text{paths from 1 to 1 in } 2n \text{ steps which stay} \geq 1\} = \binom{2n}{n} - \binom{2n}{n+1} = \frac{1}{n+1}\binom{2n}{n}.$$

(This is the nth Catalan number*, which comes up very frequently in all sorts of counting problems.)*

Hence, we have that

$$\mathbb{P}\left(S_{2n+1} = 0 \text{ and } S_k \geq 1, 1 \leq k \leq 2n | S_0 = 1\right) = \frac{1}{n+1}\binom{2n}{n} p^n q^{n+1}$$

and so, by $(*)$,

$$u_1 = \sum_{n=0}^{\infty} \frac{1}{n+1} \binom{2n}{n} p^n q^{n+1} = q \sum_{n=0}^{\infty} \frac{1}{n+1} \binom{2n}{n} (pq)^n = \frac{1 - \sqrt{1-4pq}}{2p},$$

where the last equality follows from a Taylor expansion. Now, $1 - 4pq = 1 - 4p + 4p^2 = (2p-1)^2$ and so, since $p > 1/2$, $\sqrt{1-4pq} = 2p - 1$. Hence, $u_1 = q/p$. The desired result follows. $\square$

Chapter 5

Probability generating functions

We're now going to turn to an extremely powerful tool, not just in calculations but also in proving more abstract results about discrete random variables.

From now on we consider *non-negative integer-valued* random variables i.e. X takes values in $\{0, 1, 2, \ldots\}$.

Definition 5.1. *Let X be a non-negative integer-valued random variable. The* probability generating function *(p.g.f.) of X is $G_X : \mathbb{R} \to \mathbb{R}$ defined by*

$$G_X(s) = \mathbb{E}[s^X]$$

for all $s \in \mathbb{R}$ such that the expectation exists, that is for which $\sum_{k=0}^{\infty} |s|^k \mathbb{P}(X = k) < \infty$.

Let us agree to save space by setting

$$p_k = p_X(k) = \mathbb{P}(X = k).$$

Notice that because $\sum_{k=0}^{\infty} p_k = 1$, $G_X(s)$ is certainly defined for $|s| \leq 1$ and $G_X(1) = 1$. Notice also that $G_X(s)$ is just a real-valued *function*. The parameter s is the *argument* of the function and has nothing to do with X. It plays the same role as x if I write $\sin x$, for example.[1]

Why are generating functions so useful? Because they encode all of the information about the distribution of X in a single function. It will turn out that we can get at this information by using the tools of calculus.

Theorem 5.2. *The distribution of X is uniquely determined by its probability generating function, G_X.*

Proof. First note that $G_X(0) = p_0$. Now, for $|s| < 1$, we can differentiate $G_X(s)$ term-by-term to get

$$G'_X(s) = p_1 + 2p_2 s + 3p_3 s^2 + \cdots .$$

[1] The probability generating function is an example of a *power series*, that is a function of the form $f(x) = \sum_{n=0}^{\infty} c_n x^n$. It may be that this sum is diverges for some values of x; the *radius of convergence* is the value r such that the sum converges if $|x| < r$ and diverges if $|x| > r$. For a probability generating function, we can see that the radius of convergence must be at least 1. For the purposes of this course, you are safe to assume that the derivative of f is well-defined for $|x| < r$ and is given by

$$f'(x) = \sum_{n=1}^{\infty} n c_n x^{n-1}$$

i.e. what you would get differentiating term-by-term. You will learn more about power series in Analysis I & II.

Setting $s = 0$, we see that $G'_X(0) = p_1$. Similarly, by differentiating repeatedly, we see that

$$\frac{d^k}{ds^k} G_X(s)\big|_{s=0} = k!\, p_k.$$

So we can recover $p_0, p_1, \ldots$ from G_X. □

Probability generating functions for common distributions.

1. **Bernoulli distribution.** $X \sim \mathrm{Ber}(p)$. Then

$$G_X(s) = \sum_k p_k s^k = qs^0 + ps^1 = q + ps$$

for all $s \in \mathbb{R}$.

2. **Binomial distribution.** $X \sim \mathrm{Bin}(n, p)$. Then

$$G_X(s) = \sum_{k=0}^{n} s^k \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=0}^{n} \binom{n}{k} (ps)^k (1-p)^{n-k} = (1 - p + ps)^n,$$

by the binomial theorem. This is valid for all $s \in \mathbb{R}$.

3. **Poisson distribution.** $X \sim \mathrm{Po}(\lambda)$. Then

$$G_X(s) = \sum_{k=0}^{\infty} s^k \frac{\lambda^k e^{-\lambda}}{k!} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{(s\lambda)^k}{k!} = e^{\lambda(s-1)}$$

for all $s \in \mathbb{R}$.

4. **Geometric distribution with parameter p.** Exercise on Problem Sheet 5: check that

$$G_X(s) = \frac{ps}{1 - (1-p)s},$$

provided that $|s| < \frac{1}{1-p}$.

Theorem 5.3. *If X and Y are independent, then*

$$G_{X+Y}(s) = G_X(s) G_Y(s).$$

Proof. We have

$$G_{X+Y}(s) = \mathbb{E}\left[s^{X+Y}\right] = \mathbb{E}\left[s^X s^Y\right].$$

Since X and Y are independent, s^X and s^Y are independent (by Question 2 on Problem Sheet 4). So then by Theorem 2.24, this is equal to

$$\mathbb{E}\left[s^X\right] \mathbb{E}\left[s^Y\right] = G_X(s) G_Y(s).$$ □

This can be very useful for proving distributional relationships.

Theorem 5.4. *Suppose that $X_1, X_2, \ldots, X_n$ are independent* $\mathrm{Ber}(p)$ *random variables and let $Y = X_1 + \cdots + X_n$. Then $Y \sim \mathrm{Bin}(n, p)$.*

Proof. We have

$$G_Y(s) = \mathbb{E}[s^Y] = \mathbb{E}[s^{X_1+\cdots+X_n}] = \mathbb{E}[s^{X_1}\cdots s^{X_n}] = \mathbb{E}[s^{X_1}]\cdots\mathbb{E}[s^{X_n}] = (1-p+ps)^n.$$

As Y has the same p.g.f. as a $\mathrm{Bin}(n,p)$ random variable, we deduce that $Y \sim \mathrm{Bin}(n,p)$. □

The interpretation of this is that X_i tells us whether the ith of a sequence of independent coin flips is heads or tails (where heads has probability p). Then Y counts the number of heads in n independent coin flips and so must be distributed as $\mathrm{Bin}(n,p)$.

Theorem 5.5. *Suppose that $X_1, X_2, \ldots, X_n$ are independent random variables such that $X_i \sim \mathrm{Po}(\lambda_i)$. Then*

$$\sum_{i=1}^n X_i \sim \mathrm{Po}\left(\sum_{i=1}^n \lambda_i\right).$$

In particular, if $\lambda_i = \lambda$ for all $1 \le i \le n$ then $\sum_{i=1}^n X_i \sim \mathrm{Po}(n\lambda)$.

Proof. Recall that $\mathbb{E}\left[s^{X_i}\right] = e^{\lambda_i(s-1)}$. By independence,

$$\mathbb{E}\left[s^{X_1+X_2+\ldots+X_n}\right] = \prod_{i=1}^n \mathbb{E}\left[s^{X_i}\right] = \prod_{i=1}^n e^{\lambda_i(s-1)} = \exp\left((s-1)\sum_{i=1}^n \lambda_i\right).$$

Since this is the p.g.f. of the $\mathrm{Po}(\sum_{i=1}^n \lambda_i)$ distribution and probability generating functions uniquely determine distributions, the result follows. □

5.1 Calculating expectations using probability generating functions

We've already seen that differentiating $G_X(s)$ and setting $s = 0$ gives us a way to get at the probability mass function of X. Derivatives at other points can also be useful. We have

$$G'_X(s) = \frac{d}{ds}\mathbb{E}[s^X] = \frac{d}{ds}\sum_{k=0}^\infty s^k \mathbb{P}(X=k) = \sum_{k=0}^\infty \frac{d}{ds} s^k \mathbb{P}(X=k) = \sum_{k=0}^\infty k s^{k-1}\mathbb{P}(X=k) = \mathbb{E}[Xs^{X-1}].$$

So

$$G'_X(1) = \mathbb{E}[X]$$

(as long as $\mathbb{E}[X]$ exists). Differentiating again, we get

$$G''_X(1) = \mathbb{E}[X(X-1)] = \mathbb{E}[X^2] - \mathbb{E}[X],$$

and so, in particular,

$$\mathrm{var}(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2.$$

In general,

$$\left.\frac{d^k}{ds^k} G_X(s)\right|_{s=1} = \mathbb{E}[X(X-1)\cdots(X-k+1)].$$

Example 5.6. *Let $Y = X_1 + X_2 + X_3$, where X_1, X_2 and X_3 are independent random variables each having probability generating function*

$$G(s) = \frac{1}{6} + \frac{s}{3} + \frac{s^2}{2}.$$

1. *Find the mean and variance of X_1.*
2. *What is the p.g.f. of Y? What is $\mathbb{P}(Y=3)$?*
3. *What is the p.g.f. of $3X_1$? Why is it not the same as the p.g.f. of Y? What is $\mathbb{P}(3X_1=3)$?*

Solution. 1. Differentiating the probability generating function,

$$G'(s) = \frac{1}{3} + s, \quad G''(s) = 1,$$

and so $\mathbb{E}[X_1] = G'(1) = \frac{4}{3}$ and

$$\text{var}\,(X_1) = G''(1) + G'(1) - (G'(1))^2 = 1 + \frac{4}{3} - \frac{16}{9} = \frac{5}{9}.$$

2. Just as in our derivation of the probability generating function for the binomial distribution,

$$G_Y(s) = \mathbb{E}[s^{X_1+X_2+X_3}] = \mathbb{E}[s^{X_1}]\mathbb{E}[s^{X_2}]\mathbb{E}[s^{X_3}]$$

and so

$$G_Y(s) = \left(\frac{1}{6} + \frac{s}{3} + \frac{s^2}{2}\right)^3 = \frac{1}{216}\left(1 + 6s + 21s^2 + 44s^3 + 63s^4 + 54s^5 + 27s^6\right).$$

$\mathbb{P}(Y=3)$ is the coefficient of s^3 in $G_Y(s)$, that is $\frac{11}{54}$. (As an exercise, calculate $\mathbb{P}(Y=3)$ directly.)

3. We have

$$G_{3X_1}(s) = \mathbb{E}[s^{(3X_1)}] = \mathbb{E}[(s^3)^{X_1}] = G_{X_1}(s^3) = \frac{1}{6} + \frac{s^3}{3} + \frac{s^6}{2}.$$

This is different from $G_Y(s)$ because $3X_1$ and S_3 have different distributions - knowing X_1 does not tell you Y, but it does tell you $3X_1$. Finally, $\mathbb{P}(3X_1=3) = \mathbb{P}(X_1=1) = \frac{1}{3}$. □

Of course, for each fixed $s \in \mathbb{R}$, s^X is itself a discrete random variable. So we can use the law of total probability when calculating its expectation.

Example 5.7. *Suppose that there are n red balls, n white balls and 1 blue ball in an urn. A ball is selected at random and then replaced. Let X be the number of red balls selected before a blue ball is chosen. Find*

(a) the probability generating function of X,

(b) $\mathbb{E}[X]$,

(c) $\text{var}\,(X)$.

Solution. (a) We will use the law of total probability for expectations. Let R be the event that the first ball is red, W be the event that the first ball is white and B be the event that the first ball is blue. Then

$$G_X(s) = \mathbb{E}\left[s^X\right] = \mathbb{E}\left[s^X|R\right]\mathbb{P}(R) + \mathbb{E}\left[s^X|W\right]\mathbb{P}(W) + \mathbb{E}\left[s^X|B\right]\mathbb{P}(B).$$

Of course, the value of X is affected by the first ball which is picked. If the first ball is blue then we know that $X=0$. If the first ball is white, we learn nothing about the value of X. If the first ball is red

then effectively we start over again counting numbers of red balls, but we add 1 for the red ball we have already seen. This yields

$$\begin{aligned} G_X(s) &= \mathbb{E}\left[s^{1+X}\right]\mathbb{P}(R) + \mathbb{E}\left[s^X\right]\mathbb{P}(W) + \mathbb{E}\left[s^0\right]\mathbb{P}(B) \\ &= sG_X(s)\frac{n}{2n+1} + G_X(s)\frac{n}{2n+1} + \frac{1}{2n+1} \end{aligned}$$

and so

$$G_X(s) = \frac{1}{n+1-ns} = \frac{1/(n+1)}{1-(1-1/(n+1))s}.$$

(b) Differentiating, we get

$$G'_X(s) = \frac{n}{(n+1-ns)^2}$$

and so

$$\mathbb{E}[X] = G'_X(1) = n.$$

(c) Recall that

$$\mathrm{var}(X) = G''_X(1) + G'_X(1) - (G'_X(1))^2.$$

Differentiating the p.g.f. again we get

$$G''_X(s) = \frac{2n^2}{(n+1-ns)^3}$$

and so $G''_X(1) = 2n^2$. Hence,

$$\mathrm{var}(X) = 2n^2 + n - n^2 = n(n+1).$$

If we were just asked for $\mathbb{E}[X]$ it would be easier to calculate

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[X|R]\,\mathbb{P}(R) + \mathbb{E}[X|W]\,\mathbb{P}(W) + \mathbb{E}[X|B]\,\mathbb{P}(B) \\ &= (1+\mathbb{E}[X])\frac{N}{2n+1} + \mathbb{E}[X]\,\frac{n}{2n+1} + 0\cdot\frac{1}{2n+1} = n. \end{aligned}$$

In order to calculate $\mathrm{var}(X)$, however, we need both $\mathbb{E}[X]$ and $\mathbb{E}\left[X^2\right]$ and so it's easier just to find $G_X(s)$ and differentiate it. □

Definition 5.8. *Suppose that $X_1, X_2, \ldots$ are independent random variables which all have the same distribution. Then we say that $X_1, X_2, \ldots$ are* independent and identically distributed (i.i.d.).

Theorem 5.9. *Let $X_1, X_2, \ldots$ be i.i.d. non-negative integer-valued random variables with p.g.f. $G_X(s)$. Let N be another non-negative integer-valued random variable, independent of $X_1, X_2, \ldots$ and with p.g.f. $G_N(s)$. Then the p.g.f. of $\sum_{i=1}^N X_i$ is $G_N(G_X(s))$.*

Notice that the sum $\sum_{i=1}^N X_i$ has a *random* number of terms. We interpret it as 0 if $N = 0$.

Proof. We partition according to the value of N: we have

$$\begin{aligned}
\mathbb{E}\left[s^{X_1+\cdots+X_N}\right] &= \sum_{n=0}^{\infty} \mathbb{E}\left[s^{X_1+\cdots+X_N}|N=n\right]\mathbb{P}\left(N=n\right) \quad \text{by the law of total probability} \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left[s^{X_1+\cdots+X_n}|N=n\right]\mathbb{P}\left(N=n\right) \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left[s^{X_1+\cdots+X_n}\right]\mathbb{P}\left(N=n\right) \quad \text{by the independence of } N \text{ and } \{X_1, X_2, \ldots\} \\
&= \sum_{n=0}^{\infty} \mathbb{E}\left[s^{X_1}\right]\cdots\mathbb{E}\left[s^{X_n}\right]\mathbb{P}\left(N=n\right) \quad \text{since } X_1, X_2, \ldots \text{ are independent} \\
&= \sum_{n=0}^{\infty} (G_X(s))^n \mathbb{P}\left(N=n\right) \\
&= G_N(G_X(s)).
\end{aligned}$$

□

Corollary 5.10. *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed* $\mathrm{Ber}(p)$ *random variables and that $N \sim \mathrm{Po}(\lambda)$, independently of $X_1, X_2, \ldots$. Then $\sum_{i=1}^{N} X_i \sim \mathrm{Po}(\lambda p)$.*

(Notice that we saw this result in disguise via a totally different method on Problem Sheet 4.)

Proof. We have $G_X(s) = 1 - p + ps$ and $G_N(s) = \exp(\lambda(s-1))$ and so by Theorem 5.9,

$$\mathbb{E}\left[s^{\sum_{i=1}^{N} X_i}\right] = G_N(G_X(s)) = \exp(\lambda(1-p+ps-1)) = \exp(\lambda p(s-1)).$$

Since this is the p.g.f. of $\mathrm{Po}(\lambda p)$ and p.g.f.'s uniquely determine distributions, the result follows. □

Example 5.11. *In a short fixed time period, a photomultiplier detects 0, 1 or 2 photons with probabilities $\frac{1}{2}$, $\frac{1}{3}$ and $\frac{1}{6}$ respectively. The photons detected by the photomultiplier cause it to give off a charge of 2, 3, 4 or 5 electrons (with equal probability) independently for every one photon originally detected. What is the probability generating function of the number of electrons given off in the time period? What is the probability that exactly five electrons are given off in that period?*

Solution. Let N be the number of photons detected. Then the probability generating function of N is

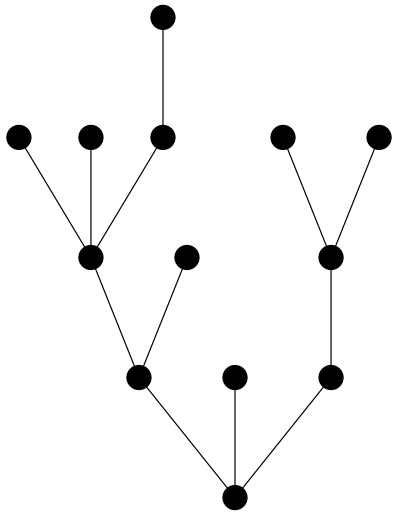
$$G_N(s) = \frac{1}{2} + \frac{1}{3}s + \frac{1}{6}s^2.$$

Let X_i be the number of electrons given off by the ith photon detected. Then $Y = X_1 + \cdots + X_N$ is the total number given off in the period (remember that N here is *random*). Now $G_X(s) = \frac{1}{4}(s^2+s^3+s^4+s^5)$ and so, by Theorem 5.9,

$$\begin{aligned}
G_Y(s) &= G_N(G_X(s)) \\
&= \frac{1}{2} + \frac{1}{3}G_X(s) + \frac{1}{6}(G_X(s))^2 \\
&= \frac{1}{2} + \frac{1}{12}s^2 + \frac{1}{12}s^3 + \frac{1}{12}s^4 + \frac{1}{12}s^5 + \frac{1}{96}(s^4 + 2s^5 + 3s^6 + 4s^7 + 3s^8 + 2s^9 + s^{10}).
\end{aligned}$$

The probability that five electrons are given off is the coefficient of s^5, that is $\frac{5}{48}$. □

Example 5.12 (Branching process)**.** *Suppose we have a population (say of bacteria). Each individual in the population gives birth to a random number of children in the next generation. This number of children has probability mass function $p(i), i \geq 0$. Each child then reproduces independently in the same manner as the parent. Here is a possible family tree of such a population:*



We start at the bottom of the tree, with a single individual in generation 0. Then there are 3 individuals in generations 1 and 2, 5 individuals in generation 3, a single individual in generation 4 and no individuals in subsequent generations. (Notice that in this case we must have $p(0) > 0$!)

Suppose that G is the probability generating function associated with $p(i), i \geq 0$ and that μ is its mean. What is the probability generating function, G_n, of the population size in generation n? What is the expected population size in generation n?

Solution. Let X_n be the size of the population in generation n. Then $X_0 = 1$ and X_1 has p.m.f. $p(i), i \geq 0$, so that $G_1(s) = \mathbb{E}\left[s_1^X\right] = G(s)$. Each individual i in generation n has a number C_i of children, for $1 \leq i \leq X_n$. So we have

$$X_{n+1} = \sum_{i=1}^{X_n} C_i,$$

where $C_1, C_2, \ldots$ are independent and identically distributed with p.g.f. G. We interpret this sum as 0 if $X_n = 0$. So

$$G_{n+1}(s) = \mathbb{E}\left[s^{X_{n+1}}\right] = \mathbb{E}\left[s^{\sum_{i=1}^{X_n} C_i}\right] = G_n(G(s)).$$

Hence, by induction, for $n \geq 1$,

$$G_n(s) = \underbrace{G(G(\ldots G}_{n \text{ times}}(s)\ldots)).$$

Now, $\mathbb{E}[X_n] = G_n'(1)$. By the chain rule,

$$G_n'(s) = \frac{d}{ds} G(G_{n-1}(s)) = G_{n-1}'(s) G'(G_{n-1}(s)).$$

Plugging in $s = 1$, we get

$$\mathbb{E}[X_n] = \mathbb{E}[X_{n-1}]\, G'(1) = \mathbb{E}[X_{n-1}]\, \mu = \cdots = \mu^n.$$

In particular, notice that we get exponential growth on average. □

Chapter 6

Random samples and the weak law of large numbers

Definition 6.1. *Let $X_1, X_2, \ldots, X_n$ denote i.i.d. random variables. Then these random variables are said to constitute a* random sample *of size n from the distribution.*

Statistics often involves random samples where the underlying distribution (the *"parent distribution"*) is unknown. A realisation of such a random sample is used to make inferences about the parent distribution. Suppose, for example, we want to know about the mean of the parent distribution. An important estimator is the sample mean.

Definition 6.2. *The* sample mean *is defined to be* $\bar{X}_n = \dfrac{1}{n}\sum_{i=1}^{n} X_i$.

This is a key random variable which itself has an expectation and a variance. Recall that for discrete random variables X and Y,

$$\operatorname{var}(X+Y) = \operatorname{var}(X) + \operatorname{var}(Y) + 2\operatorname{cov}(X,Y).$$

We can extend this (by induction) to n random variables as follows:

$$\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \operatorname{var}(X_i) + \sum_{i\neq j} \operatorname{cov}(X_i, X_j).$$

Theorem 6.3. *Suppose that $X_1, X_2, \ldots, X_n$ form a random sample from a distribution with mean μ and variance σ^2. Then the expectation and variance of the sample mean are*

$$\mathbb{E}\left[\bar{X}_n\right] = \mu \quad \textit{and} \quad \operatorname{var}\left(\bar{X}_n\right) = \frac{1}{n}\sigma^2.$$

Proof. We have $\mathbb{E}[X_i] = \mu$ and $\operatorname{var}(X_i) = \sigma^2$ for $1 \leq i \leq n$. So

$$\mathbb{E}\left[\bar{X}_n\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] = \mu,$$

$$\operatorname{var}\left(\bar{X}_n\right) = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\operatorname{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} \operatorname{var}(X_i) = \frac{1}{n}\sigma^2,$$

since independence implies that $\operatorname{cov}(X_i, X_j) = 0$ for all $i \neq j$. □

Example 6.4. *Let $X_1, \ldots, X_n$ be a random sample from a Bernoulli distribution with parameter p. Then $\mathbb{E}[X_i] = p$, $\operatorname{var}(X_i) = p(1-p)$ for all $1 \leq i \leq n$. Hence, $\mathbb{E}\left[\bar{X}_n\right] = p$ and $\operatorname{var}\left(\bar{X}_n\right) = p(1-p)/n$.*

In order for $\bar{X}_n$ to be a good estimator of the mean, we would like to know that for large sample sizes n, $\bar{X}_n$ is not too far away from μ i.e. that $|\bar{X}_n - \mu|$ is small. The result which tells us that this is true is called the *law of large numbers* and is of fundamental importance in probability. Before we state it, let's step away from the sample mean and consider a more basic situation.

Suppose that A is an event with probability $\mathbb{P}(A)$ and write $p = \mathbb{P}(A)$. Let X be the indicator function of the event A i.e. the random variable defined by

$$X(\omega) = \mathbb{1}_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{if } \omega \notin A. \end{cases}$$

Then $X \sim \text{Ber}(p)$ and $\mathbb{E}[X] = p$. Suppose now that we perform our experiment repeatedly and let X_i be the indicator of the event that A occurs on the ith trial. Our intuitive notion of probability leads us to believe that if the number n of trials is large then the *proportion* of the time that A occurs should be close to p i.e.

$$\left| \frac{1}{n} \sum_{i=1}^{n} X_i - p \right|$$

should be small. So proving that the sample mean is close to the true mean in this situation will also provide some justification for the way we have set up our mathematical theory of probability.

Theorem 6.5 (Weak law of large numbers)**.** *Suppose that $X_1, X_2, \ldots$ are independent and identically distributed random variables with mean μ and variance $\sigma^2 < \infty$. Then for any fixed $\epsilon > 0$,*

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| > \epsilon \right) \to 0$$

as $n \to \infty$.

(Equivalently, we could have put

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} X_i - \mu \right| \leq \epsilon \right) \to 1$$

as $n \to \infty$.)

In other words, the probability that the sample mean deviates from the true mean by more than some small quantity ϵ tends to 0 as $n \to \infty$. Notice that the result only depends on the underlying distribution through its mean. There is also a more general version which does not require the finiteness of the variance, but we won't discuss it here.

In order to prove the weak law, we need a very useful inequality.

Theorem 6.6 (Chebyshev's inequality)**.** *Suppose that Y is a random variable such that $\mathbb{E}\left[Y^2\right]$ exists. Then for any $t > 0$,*

$$\mathbb{P}(|Y| > t) \leq \frac{\mathbb{E}\left[Y^2\right]}{t^2}.$$

Proof. Let $A = \{|Y| > t\}$. We may assume that $\mathbb{P}(A) > 0$, since otherwise the result is trivially true. Then by the law of total probability for expectations,

$$\mathbb{E}\left[Y^2\right] = \mathbb{E}\left[Y^2|A\right]\mathbb{P}(A) + \mathbb{E}\left[Y^2|A^c\right]\mathbb{P}(A^c) \geq \mathbb{E}\left[Y^2|A\right]\mathbb{P}(A),$$

since $\mathbb{P}(A^c) \geq 0$ and $\mathbb{E}\left[Y^2|A^c\right] \geq 0$. Now, we certainly have $\mathbb{E}\left[Y^2|A\right] = \mathbb{E}\left[Y^2||Y| > t\right] > t^2$. So, rearranging, we get

$$\mathbb{P}(|Y| > t) \leq \frac{\mathbb{E}\left[Y^2\right]}{t^2}$$

as we wanted. □

Proof of Theorem 6.5. Set

$$Y = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu.$$

Then

$$\mathbb{E}\left[Y^2\right] = \mathbb{E}\left[\left(\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right)^2\right] = \operatorname{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{\sigma^2}{n}.$$

So by Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu\right| > \epsilon\right) \leq \frac{\sigma^2}{n\epsilon^2}.$$

Since $\epsilon > 0$ is fixed, the right-hand side tends to 0 as $n \to \infty$. □

Chapter 7

Continuous random variables

7.1 Random variables and cumulative distribution functions

Recall that we defined a discrete random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to be a function $X : \Omega \to \mathbb{R}$ such that X can only take countably many values (and such that we can assign a probability to the event $\{X = x\}$, i.e. such that $\{X = x\} \in \mathcal{F}$). There is, however, a more general notion. The essential idea is that a random variable can be *any* (sufficiently nice) function $X : \Omega \to \mathbb{R}$, which represents some sort of observable quantity in our random experiment.

Why do we need more general random variables?

- Some outcomes are essentially continuous. In particular, many physical quantities are most naturally modelled as taking uncountably many possible values, for example, lengths, weights and speeds.
- Even for discrete quantities, it is often useful to think instead in terms of continuous approximations. For example, suppose you wish to calculate the number of working adults who regularly contribute to charity. You might model this number as X out of n, where n is the total number of working adults in the UK. We could, in theory, model this as a $\text{Bin}(n, p)$ random variable where $p = \mathbb{P}(\text{adult contributes})$. But n is measured in millions. So instead model $Y \approx \frac{X}{n}$ as a continuous random variable taking values in $[0, 1]$ and giving the proportion of adults who contribute.

To give a concrete example of a random variable which is not discrete, imagine you have a board game spinner. You spin the arrow and it lands pointing at an angle somewhere between 0 and 2π in such a way that every angle is equally likely; we want to model this angle as a random variable X. How can we describe its distribution? We can't assign a positive probability to each angle – our probabilities wouldn't sum to 1. To get around this, we don't define the probability of individual sample points, but only of certain natural events. For example, by symmetry we expect that $\mathbb{P}(X \leq \pi) = 1/2$. More generally, we expect the probability that X lies in an interval $[a, b] \subseteq [0, 2\pi)$ to be proportional to the length of that interval: $\mathbb{P}(X \in [a, b]) = \frac{b-a}{2\pi}$, $0 \leq a < b < 2\pi$.

Definition 7.1. *A* random variable X *defined on a probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is a function* $X : \Omega \to \mathbb{R}$ *such that* $\{\omega : X(\omega) \leq x\} \in \mathcal{F}$ *for each* $x \in \mathbb{R}$.

Let's just check that this includes our earlier definition. If X is a discrete random variable then

$$\{\omega : X(\omega) \leq x\} = \bigcup_{y \leq x : y \in \mathrm{Im}X} \{\omega : X(\omega) = y\}.$$

Since $\mathrm{Im}X$ is countable, this is a countable union of events in $\mathcal{F}$ and, therefore, itself belongs to $\mathcal{F}$.

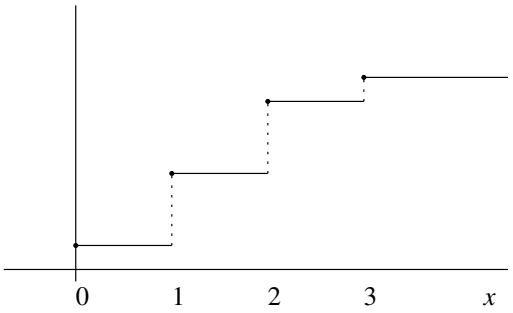
Of course, $\{\omega : X(\omega) \leq x\} \in \mathcal{F}$ means precisely that we can assign a probability to this event. The collection of these probabilities as x varies in $\mathbb{R}$ will play a central part in what follows.

Definition 7.2. *The* cumulative distribution function (c.d.f.) *of a random variable X is the function $F_X : \mathbb{R} \to [0,1]$ defined by*

$$F_X(x) = \mathbb{P}(X \leq x).$$

Example 7.3. *Let X be the number of heads obtained in three tosses of a fair coin. Then $\mathbb{P}(X = 0) = \frac{1}{8}$, $\mathbb{P}(X = 1) = \mathbb{P}(X = 2) = \frac{3}{8}$ and $\mathbb{P}(X = 3) = \frac{1}{8}$. So*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0 \\ \frac{1}{8} & \text{if } 0 \leq x < 1 \\ \frac{1}{8} + \frac{3}{8} = \frac{1}{2} & \text{if } 1 \leq x < 2 \\ \frac{1}{8} + \frac{3}{8} + \frac{3}{8} = \frac{7}{8} & \text{if } 2 \leq x < 3 \\ 1 & \text{if } x \geq 3. \end{cases}$$



Example 7.4. *Let X be the angle of the board game spinner. Then*

$$F_X(x) = \begin{cases} 0 & \text{if } x < 0, \\ \frac{x}{2\pi} & \text{if } 0 \leq x < 2\pi, \\ 1 & \text{if } x \geq 2\pi. \end{cases}$$

We can immediately write down some properties of the c.d.f. F_X corresponding to a general random variable X.

Theorem 7.5. *1. F_X is non-decreasing.*

2. $\mathbb{P}(a < X \leq b) = F_X(b) - F_X(a)$ for $a < b$.

3. As $x \to -\infty$, $F_X(x) \to 0$.

4. As $x \to \infty$, $F_X(x) \to 1$.

Proof. 1. If $a < b$ then $\{\omega : X(\omega) \leq a\} \subseteq \{\omega : X(\omega) \leq b\}$ and so

$$F_X(a) = \mathbb{P}(X \leq a) \leq \mathbb{P}(X \leq b) = F_X(b).$$

2. Since $\{X \leq a\}$ is a subset of $\{X \leq b\}$,

$$\mathbb{P}(a < X \leq b) = \mathbb{P}(\{X \leq b\} \setminus \{X \leq a\}) = \mathbb{P}(X \leq b) - \mathbb{P}(X \leq a) = F_X(b) - F_X(a).$$

3 & 4. (sketch) Intuitively, we want to put "$F_X(-\infty) = \mathbb{P}(X \leq -\infty)$" and then, since X can't possibly be $-\infty$ (or less!), the only sensible interpretation we could give the right-hand side would be 0. Likewise,

we would like to put "$F_X(\infty) = \mathbb{P}(X \leq \infty)$" and, since X cannot be larger than ∞, the only sensible interpretation we could give the right-hand side would be 1. The problem is that ∞ and $-\infty$ aren't real numbers, but F_X is a function on $\mathbb{R}$. The only sensible way to deal with this problem is by taking limits instead.[1] □

Conversely, any function F satisfying conditions 1, 3 and 4 of Theorem 7.5 plus *right-continuity* is the cumulative distribution function of *some* random variable defined on *some* probability space, although we will not prove this fact.

As you can see from the coin-tossing example, F_X need not be a smooth function. Indeed, for a discrete random variable, F_X is always a step function. However, in the rest of the course, we're going to concentrate on the case where F_X has a derivative (i.e. F_X is very smooth).

Definition 7.6. *A* continuous random variable X *is a random variable whose c.d.f. satisfies*

$$F_X(x) = \mathbb{P}(X \leq x) = \int_{-\infty}^{x} f_X(u)du,$$

where $f_X : \mathbb{R} \to \mathbb{R}$ *is a function such that*

(a) $f_X(u) \geq 0$ *for all* $u \in \mathbb{R}$

(b) $\int_{-\infty}^{\infty} f_X(u)du = 1$.

By the Fundamental Theorem of Calculus, we then have that $F_X(x)$ is differentiable with

$$\frac{dF_X(x)}{dx} = f_X(x).$$

f_X is called the *probability density function (p.d.f.)* of X or, sometimes, just its *density*.

Example 7.7. *Suppose that a continuous random variable* X *has p.d.f.*

$$f_X(x) = \begin{cases} cx^2(1-x) & \text{for } x \in [0,1] \\ 0 & \text{otherwise.} \end{cases}$$

Find the constant c *and an expression for the c.d.f.*

[1] The following rigorous proof of 3 and 4 is here just for interest, and is non-examinable. You are welcome to ignore it! Recall the alternative version of axiom $\mathbf{P}_4$:

$\mathbf{P}_4'$: If $A_1 \supseteq A_2 \supseteq \ldots$ is a sequence from $\mathcal{F}$ with $\cap_n A_n = \emptyset$, then $(\mathbb{P}(A_n))_{n\geq 1}$ is a decreasing sequence which tends to 0 as $n \to \infty$.

3. First notice that since $F_X(x)$ decreases as x decreases, it's sufficient to show that $F_X(-n) \to 0$ as $n \to \infty$ through the integers. Take $A_n = \{\omega : X(\omega) \leq -n\}$ so that $\mathbb{P}(A_n) = F_X(-n)$. Then, $A_n \supseteq A_{n+1}$ for all $n \geq 1$. Since no real number is smaller than all of the negative integers, we have

$$\bigcap_{i\geq 1} A_i = \emptyset.$$

Hence by $\mathbf{P}_4'$, $F_X(-n) \to 0$ as $n \to \infty$.

4. Again, since $F_X(x)$ increases as x increases, it's sufficient to show that $F_X(n) \to 1$ as $n \to \infty$. By taking complements in $\mathbf{P}_4'$, we get that if $B_1 \subseteq B_2 \subseteq \ldots$ are events with $\cup_{i\geq 1} B_i = \Omega$ then

$$\mathbb{P}(B_n) \to 1$$

as $n \to \infty$. Take $B_n = \{\omega : X(\omega) \leq n\}$ so that $\mathbb{P}(B_n) = F_X(n)$ and $B_n \subseteq B_{n+1}$ for all $n \geq 1$. Moreover, since every real number is smaller than some positive integer,

$$\bigcup_{i\geq 1} B_i = \Omega.$$

The result follows.

Solution. To find the constant, c, note that we must have

$$1 = \int_{-\infty}^{\infty} f_X(x)dx = \int_0^1 cx^2(1-x)dx = c\left[\frac{x^3}{3} - \frac{x^4}{4}\right]_0^1 = \frac{c}{12}.$$

It follows that $c = 12$. To find the c.d.f., we simply integrate:

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx = \begin{cases} 0 & \text{for } x < 0 \\ \int_0^x 12x^2(1-x)dx & \text{for } 0 \leqslant x < 1 \\ 1 & \text{for } x \geqslant 1. \end{cases}$$

Since

$$\int_0^x 12x^2(1-x)dx = 12\left(\frac{x^3}{3} - \frac{x^4}{4}\right),$$

we get

$$F_X(x) = \begin{cases} 0 & \text{for } x < 0 \\ 4x^3 - 3x^4 & \text{for } 0 \leqslant x < 1 \\ 1 & \text{for } x \geqslant 1. \end{cases}$$

□

Example 7.8. *The duration in minutes of mobile phone calls made by students is modelled by a random variable, X, with p.d.f.*

$$f_X(x) = \begin{cases} \frac{1}{6}e^{-x/6} & \text{if } x \geq 0 \\ 0 & \text{otherwise.} \end{cases}$$

What is the probability that a call lasts

(i) between 3 and 6 minutes?

(ii) more than 6 minutes?

Solution. (i)

$$\mathbb{P}(3 < X \leq 6) = \int_3^6 f_X(x)\mathrm{d}x = \int_3^6 \frac{1}{6}e^{-x/6}dx = e^{-\frac{1}{2}} - e^{-1}.$$

(ii)

$$\mathbb{P}(X > 6) = \int_6^{\infty} f_X(x)\mathrm{d}x = \int_6^{\infty} \frac{1}{6}e^{-x/6}dx = e^{-1}.$$

□

We often use the p.d.f. of a continuous random variable analogously to the way we used the p.m.f. of a discrete random variable. There are several similarities between the two:

Probability density function (continuous)	Probability mass function (discrete)
$f_X(x) \geqslant 0 \quad \forall x \in \mathbb{R}$	$p_X(x) \geqslant 0 \quad \forall x \in \mathbb{R}$
$\int_{-\infty}^{\infty} f_X(x) = 1$	$\sum_{x \in \mathrm{Im}X} p_X(x) = 1$
$F_X(x) = \int_{-\infty}^{x} f_X(u)du$	$F_X(x) = \sum_{u \leqslant x:\ u \in \mathrm{Im}X} p_X(u)$

However, the analogy can be misleading. For example, there's nothing to prevent $f_X(x)$ exceeding 1.

WARNING: $f_X(x)$ IS NOT A PROBABILITY.

Suppose that $\epsilon > 0$ is small. Then, by Taylor's theorem,

$$\mathbb{P}(x < X \leq x + \epsilon) = F_X(x + \epsilon) - F_X(x) \approx f_X(x)\epsilon.$$

So $f_X(x)\epsilon$ is approximately the probability that X falls between x and $x + \epsilon$ (or, indeed, between $x - \epsilon$ and x). What happens as $\epsilon \to 0$?

Theorem 7.9. *If X is a continuous random variable with p.d.f. f_X then*

$$\mathbb{P}(X = x) = 0 \quad \textit{for all } x \in \mathbb{R}$$

and

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx.$$

Proof. (Non-examinable.) We argue by contradiction. Suppose that for some $x \in \mathbb{R}$ we have $\mathbb{P}(X = x) > 0$. Let $p = \mathbb{P}(X = x)$. Then for all $n \geq 1$, $\mathbb{P}(x - 1/n < X \leq x) \geq p$. We have $\mathbb{P}(x - 1/n < X \leq x) = F_X(x) - F_X(x - 1/n)$ and so $F_X(x) - F_X(x - 1/n) \geq p$ for all $n \geq 1$. But F_X is continuous at x and so

$$\lim_{n\to\infty} (F_X(x) - F_X(x - 1/n)) = 0.$$

This gives a contradiction. So we must have $\mathbb{P}(X = x) = 0$.

Finally, $\mathbb{P}(a \leq X \leq b) = \mathbb{P}(X = a) + \mathbb{P}(a < X \leq b)$ and so, since $\mathbb{P}(X = a) = 0$, we get

$$\mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)dx. \qquad \square$$

So for a continuous r.v. X, the probability of getting any fixed value x is 0! Why doesn't this break our theory of probability? We have

$$\{\omega : X(\omega) \leq x\} = \bigcup_{y \leq x} \{\omega : X(\omega) = y\}$$

and the right-hand side is an *uncountable* union of disjoint events of probability 0. If the union were countable, this would entail that the left-hand side had probability 0 also, which wouldn't make much sense. But because the union is uncountable, we cannot expect to "sum up" these zeros in order to get the probability of the left-hand side. The right way to resolve this problem is using a probability density function.

Remark 7.10. *There do exist random variables which are neither discrete nor continuous. To give a slightly artificial example, if X is a discrete random variable and Y is a continuous random variable then $X + Y$ is a random variable which can take uncountably many values but does not have a density. The theory is particularly nice in the discrete and continuous cases because we can work with probability mass functions and probability density functions respectively. But the cumulative distribution function is a more general concept which makes sense for* all *random variables.*

7.2 Some classical distributions

As we did for discrete distributions, we introduce a stock of examples of continuous distributions which will come up time and again in this course.

1. **The uniform distribution.** X has the uniform distribution on an interval $[a, b]$ if it has p.d.f.

$$f_X(x) = \begin{cases} \frac{1}{b-a} & \text{for } a \leq x \leq b, \\ 0 & \text{otherwise.} \end{cases}$$

   We write $X \sim \mathrm{U}[a, b]$.

2. **The exponential distribution.** X has the exponential distribution with parameter $\lambda \geq 0$ if it has p.d.f.

$$f_X(x) = \lambda e^{-\lambda x}, \quad x \geq 0.$$

   We write $X \sim \mathrm{Exp}(\lambda)$. The exponential distribution is often used to model lifetimes or the time elapsing between unpredictable events (such as telephone calls, arrivals of buses, earthquakes, emissions of radioactive particles, etc).

3. **The gamma distribution.** X has the gamma distribution with parameters $\alpha > 0$ and $\lambda \geq 0$ if it has p.d.f.

$$f_X(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, \quad x \geq 0.$$

   Here, $\Gamma(\alpha)$ is the so-called *gamma function*, which is defined by

$$\Gamma(\alpha) = \int_0^\infty u^{\alpha-1} e^{-u} du$$

   for $\alpha > 0$. For most values of α this integral does not have a closed form. However, for a strictly positive integer n, we have $\Gamma(n) = (n-1)!$. (See the Wikipedia "Gamma function" page for lots more information about this fascinating function!)

   If X has the above p.d.f. we write $X \sim \mathrm{Gamma}(\alpha, \lambda)$. The gamma distribution is a generalisation of the exponential distribution and possesses many nice properties. The *Chi-squared distribution with d degrees of freedom*, χ^2_d, which you may have seen at 'A' Level, is the same as $\mathrm{Gamma}(d/2, 1/2)$ for $d \in \mathbb{N}$.

4. **The normal (or Gaussian) distribution.** X has the normal distribution with parameters $\mu \geq 0$ and $\sigma^2 \geq 0$ if it has p.d.f.

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right), \quad x \in \mathbb{R}.$$

   We write $X \sim \mathrm{N}(\mu, \sigma^2)$. The *standard normal distribution* is $\mathrm{N}(0, 1)$. The normal distribution is used to model all sorts of characteristics of large populations and samples. Its fundamental importance across Probability and Statistics is a consequence of the Central Limit Theorem, which you will use in Prelims Statistics and see proved in Part A Probability.

Exercise 7.11. *For the uniform and exponential distributions:*

- *Check that for each of these f_X really is a p.d.f. (i.e. that it is non-negative and integrates to 1).*
- *Calculate the corresponding c.d.f.'s.*

Example 7.12. *Show that*

$$I := \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = 1.$$

Solution. We first change variables in the integral. Set $z = (x-\mu)/\sigma$. Then

$$I = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right) dz.$$

It follows that

$$\begin{aligned} I^2 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y^2}{2}\right) dy \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{(x^2+y^2)}{2}\right) dxdy. \end{aligned}$$

Now convert to polar co-ordinates: let r and θ be such that $x = r\cos\theta$ and $y = r\sin\theta$. Then the Jacobian is $|J| = r$ and so we get

$$\int_0^{2\pi} \int_0^{\infty} \frac{1}{2\pi} r \exp\left(-\frac{r^2}{2}\right) drd\theta = \left[-e^{-r^2/2}\right]_0^{\infty} = 1.$$

Since I is clearly non-negative (it's the integral of a non-negative function), we must have $I = 1$. □

The c.d.f. of the standard normal distribution,

$$F_X(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-u^2/2} du,$$

cannot be written in a closed form, but can be found by numerical integration to an arbitrary degree of accuracy. This very important function is usually called Φ and if you did some Statistics at 'A' Level you will certainly have come across tables of its values.

7.3 Expectation

Recall that for a discrete r.v. we defined

$$\mathbb{E}[X] = \sum_{x \in \mathrm{Im}X} x p_X(x)$$

whenever the sum is absolutely convergent and, more generally, for any function $h : \mathbb{R} \to \mathbb{R}$, we had

$$\mathbb{E}[h(X)] = \sum_{x \in \mathrm{Im}X} h(x) p_X(x)$$

whenever this sum is absolutely convergent. We want to make an analogous definition for continuous random variables. Suppose X has p.d.f. f_X. Then for any x and small $\delta > 0$,

$$\mathbb{P}(x \le X \le x + \delta) \approx f_X(x)\delta$$

and, in particular,

$$\mathbb{P}(n\delta \le X \le (n+1)\delta) \approx f_X(n\delta)\delta.$$

So for the expectation, we want something like

$$\sum_{n=-\infty}^{\infty} (n\delta) f_X(n\delta)\delta.$$

We now want to take $\delta \to 0$; intuitively, we should obtain an integral.

Definition 7.13. *Let X be a continuous random variable with probability density function f_X. The* expectation *or* mean *of X is defined to be*

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

whenever $\int_{-\infty}^{\infty} |x| f_X(x) dx < \infty$. Otherwise, we say that the mean is undefined.

More generally, if $h : \mathbb{R} \to \mathbb{R}$ is any function such that $\int_{-\infty}^{\infty} |h(x)| f_X(x) dx < \infty$ then we define the expectation *of $h(X)$ to be*

$$\mathbb{E}[h(X)] = \int_{-\infty}^{\infty} h(x) f_X(x) dx.$$

As before, we define the *variance* of X to be

$$\mathrm{var}(X) = \mathbb{E}\left[(X - \mathbb{E}[X])^2\right].$$

For simplicity of notation, write $\mu = \mathbb{E}[X]$. Then we have

$$\begin{aligned}
\mathrm{var}(X) &= \int_{-\infty}^{\infty} (x-\mu)^2 f_X(x) dx \\
&= \int_{-\infty}^{\infty} (x^2 - 2x\mu + \mu^2) f_X(x) dx \\
&= \int_{-\infty}^{\infty} x^2 f_X(x) dx - 2\mu \int_{-\infty}^{\infty} x f_X(x) dx + \mu^2 \int_{-\infty}^{\infty} f_X(x) dx \\
&= \mathbb{E}\left[X^2\right] - \mu^2,
\end{aligned}$$

since $\int_{-\infty}^{\infty} x f_X(x) dx = \mu$ and $\int_{-\infty}^{\infty} f_X(x) dx = 1$. So we recover the expression

$$\mathrm{var}(X) = \mathbb{E}\left[X^2\right] - (\mathbb{E}[X])^2.$$

Just as in the discrete case, expectation has a *linearity property.*

Theorem 7.14. *Suppose X is a continuous random variable with p.d.f. f_X. Then if $a, b \in \mathbb{R}$ then $\mathbb{E}[aX + b] = a\mathbb{E}[X] + b$ and $\mathrm{var}(aX + b) = a^2 \mathrm{var}(X)$.*

Proof. We have

$$\mathbb{E}[aX + b] = \int_{-\infty}^{\infty} (ax + b) f_X(x) dx = a \int_{-\infty}^{\infty} x f_X(x) dx + b \int_{-\infty}^{\infty} f_X(x) dx = a\mathbb{E}[X] + b,$$

as required, since the density integrates to 1. Moreover,

$$\mathrm{var}(aX + b) = \mathbb{E}\left[(aX + b - a\mathbb{E}[X] - b)^2\right] = \mathbb{E}\left[a^2 (X - \mathbb{E}[X])^2\right] = a^2 \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] = a^2 \mathrm{var}(X).$$

□

Example 7.15. *Suppose $X \sim \mathrm{N}(\mu, \sigma^2)$. Then*

- *X has the same distribution as $\mu + \sigma Z$, where $Z \sim \mathrm{N}(0,1)$.*
- *X has c.d.f. $F_X(x) = \Phi((x-\mu)/\sigma)$, where Φ is the standard normal c.d.f.*
- $\mathbb{E}[X] = \mu$.
- $\mathrm{var}(X) = \sigma^2$.

Solution. First suppose that $\mu = 0$ and $\sigma^2 = 1$. Then the first two assertions are trivial and

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} e^{-x^2/2} dx$$

which must equal 0 since the integrand is an odd function. Since the mean is 0,

$$\mathrm{var}(X) = \mathbb{E}\left[X^2\right] = \int_{-\infty}^{\infty} \frac{x^2}{\sqrt{2\pi}} e^{-x^2/2} dx = \int_{-\infty}^{\infty} x \cdot \frac{x e^{-x^2/2}}{\sqrt{2\pi}} dx.$$

Integrating by parts, we get that this equals

$$\left[-x \cdot \frac{e^{-x^2/2}}{\sqrt{2\pi}}\right]_{-\infty}^{\infty} + \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx = 1.$$

So $\mathrm{var}(X) = 1$.

Suppose now that $Z \sim \mathrm{N}(0,1)$. Then

$$\mathbb{P}(\mu + \sigma Z \le x) = \mathbb{P}(Z \le (x-\mu)/\sigma) = \Phi((x-\mu)/\sigma).$$

Let $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$, the standard normal density. Differentiating $\mathbb{P}(\mu + \sigma Z \le x)$ in x, we get

$$\frac{1}{\sigma}\phi((x-\mu)/\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right).$$

So $\mu + \sigma Z \sim \mathrm{N}(\mu, \sigma^2)$. Finally,

$$\mathbb{E}[X] = \mathbb{E}[\mu + \sigma Z] = \mu + \sigma \mathbb{E}[Z] = \mu$$

and

$$\mathrm{var}(X) = \mathrm{var}(\mu + \sigma Z) = \sigma^2 \mathrm{var}(Z) = \sigma^2.$$ □

Exercise 7.16. *Show that if $X \sim \mathrm{U}[a,b]$ and $Y \sim \mathrm{Exp}(\lambda)$ then*

$$\mathbb{E}[X] = \frac{a+b}{2}, \quad \mathrm{var}(X) = \frac{(b-a)^2}{12}, \quad \mathbb{E}[Y] = \frac{1}{\lambda}, \quad \mathrm{var}(Y) = \frac{1}{\lambda^2}.$$

Notice, in particular, that the parameter of the Exponential distribution is the reciprocal *of its mean.*

Example 7.17. *Suppose that $X \sim$ Gamma$(2,2)$, so that it has p.d.f.*

$$f_X(x) = \begin{cases} 4xe^{-2x} & \text{for } x \ge 0, \\ 0 & \text{otherwise.} \end{cases}$$

Find $\mathbb{E}[X]$ and $\mathbb{E}\left[\frac{1}{X}\right]$.

Solution. We have

$$\mathbb{E}[X] = \int_{-\infty}^{\infty} x \cdot 4xe^{-2x}dx = \int_{-\infty}^{\infty} \frac{2^3}{2!}x^{3-1}e^{-2x}dx$$

and, since $\Gamma(3) = 2!$ we recognise the integrand as the density of a Gamma$(3, 2)$ random variable. So it must integrate to 1 and we get $\mathbb{E}[X] = 1$.

On the other hand,

$$\mathbb{E}\left[\frac{1}{X}\right] = \int_{-\infty}^{\infty} \frac{1}{x} \cdot 4xe^{-2x}dx = 2\int_{-\infty}^{\infty} 2e^{-2x}dx$$

and again we recognise the integrand as the density of an Exp(2) random variable which must integrate to 1. So we get $\mathbb{E}\left[\frac{1}{X}\right] = 2$.

Note that, in general, $\mathbb{E}\left[\frac{1}{X}\right] \neq \frac{1}{\mathbb{E}[X]}$. □

7.4 Functions of continuous random variables

Example 7.18. *Imagine a forest. Suppose that R is the distance from a tree to the nearest neighbouring tree. Suppose that R has p.d.f.*

$$f_R(r) = \begin{cases} re^{-r^2/2} & \text{for } r \geq 0, \\ 0 & \text{otherwise.} \end{cases}$$

Find the distribution of the tree-free area around the original tree.

Solution. Let A be the tree-free area; then $A = \pi R^2$. We begin by finding the c.d.f. of R and then use it to find the c.d.f. of A. $F_R(r)$ is clearly 0 for $r < 0$. For $r \geq 0$,

$$F_R(r) = \mathbb{P}(R \leq r) = \int_0^r se^{-s^2/2}ds = \left[-e^{-s^2/2}\right]_0^r = 1 - e^{-r^2/2}.$$

Hence, using the fact that R can't take negative values,

$$F_A(a) = \mathbb{P}(A \leq a) = \mathbb{P}\left(\pi R^2 \leq a\right) = \mathbb{P}\left(R \leq \sqrt{\frac{a}{\pi}}\right) = F_R\left(\sqrt{\frac{a}{\pi}}\right) = 1 - e^{-a/(2\pi)}$$

for $a \geq 0$. Of course, $F_A(a) = 0$ for $a < 0$. Differentiating for $a \geq 0$, we get

$$f_A(a) = \frac{1}{2\pi}e^{-a/(2\pi)}.$$

So, recognising the p.d.f., we see that A is distributed exponentially with parameter $1/(2\pi)$. □

We can generalise the idea in this example to prove the following theorem.

Theorem 7.19. *Suppose that X is a continuous random variable with density f_X and that $h : \mathbb{R} \to \mathbb{R}$ is a differentiable function which is strictly increasing (i.e. $\frac{dh(x)}{dx} > 0$ for all x). Then $Y = h(X)$ is a continuous random variable with p.d.f.*

$$f_Y(y) = f_X(h^{-1}(y))\frac{d}{dy}h^{-1}(y),$$

where h^{-1} is the inverse function of h.

Proof. Since h is strictly increasing, $h(X) \leq y$ if and only if $X \leq h^{-1}(y)$. So the c.d.f. of Y is

$$F_Y(y) = \mathbb{P}\left(h(X) \leq y\right) = \mathbb{P}\left(X \leq h^{-1}(y)\right) = F_X(h^{-1}(y)).$$

Differentiating with respect to y using the chain rule, we get

$$f_Y(y) = f_X(h^{-1}(y))\frac{d}{dy}h^{-1}(y). \qquad \square$$

There is a similar result in the case where h is strictly decreasing. In any case, you may find it easier to remember the proof than the statement of the theorem!

What if the function h is not one-to-one? It's best to treat these on a case-by-case basis and think them through carefully. Here's an example.

Example 7.20. *Suppose that a point is chosen uniformly from the perimeter of the unit circle. What is the distribution of its x-co-ordinate?*

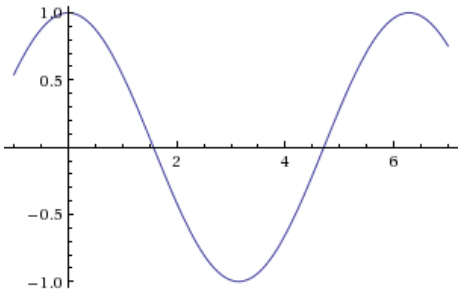
Solution. Represent the chosen point by its angle, Θ. So then Θ has a uniform distribution on $[0, 2\pi)$, with p.d.f.

$$f_\Theta(\theta) = \begin{cases} \frac{1}{2\pi} & \text{for } 0 \leq \theta < 2\pi \\ 0 & \text{otherwise.} \end{cases}$$

Moreover, the x-co-ordinate is $X = \cos\Theta$, which takes values in $[-1, 1]$. We again work via c.d.f.'s:

$$F_\Theta(\theta) = \begin{cases} 0 & \text{for } \theta < 0 \\ \frac{\theta}{2\pi} & \text{for } 0 \leq \theta < 2\pi \\ 1 & \text{for } \theta \geq 2\pi. \end{cases}$$

Notice that there are two angles in $[0, 2\pi)$ corresponding to each x-co-ordinate in $(-1, 1)$:



Then

$$\begin{aligned} F_X(x) &= \mathbb{P}\left(\cos\Theta \leq x\right) \\ &= \mathbb{P}\left(\arccos x \leq \Theta \leq 2\pi - \arccos x\right) \\ &= F_\Theta(2\pi - \arccos x) - F_\Theta(\arccos x) \\ &= 1 - \frac{\arccos x}{2\pi} - \frac{\arccos x}{2\pi} \\ &= 1 - \frac{1}{\pi}\arccos x. \end{aligned}$$

Fix $\arccos x \in [0, \pi)$. Differentiating, we get

$$f_X(x) = \begin{cases} \frac{1}{\pi} \frac{1}{\sqrt{1-x^2}} & \text{for } -1 < x < 1 \\ 0 & \text{for } x < -1 \text{ or } x > 1 \\ \text{undefined} & \text{for } x = -1 \text{ or } x = 1. \end{cases}$$

Notice that $f_X(x) \to \infty$ as $x \to 1$ or $x \to -1$ even though $\int_{-\infty}^{\infty} f_X(x)dx = 1$. □

7.5 Joint distributions

We will often want to think of different random variables defined on the same probability space. In the discrete case, we studied pairs of random variables via their joint probability mass function. For a pair of arbitrary random variables, we use instead the *joint cumulative distribution function*, $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$, given by

$$F_{X,Y}(x, y) = \mathbb{P}(X \leq x, Y \leq y).$$

It's again possible to show that this function is non-decreasing in each of its arguments, and that

$$\lim_{x \to -\infty} \lim_{y \to -\infty} F_{X,Y}(x, y) = 0$$

and

$$\lim_{x \to \infty} \lim_{y \to \infty} F_{X,Y}(x, y) = 1.$$

Definition 7.21. *Let X and Y be random variables such that*

$$F_{X,Y}(x, y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f_{X,Y}(u, v)dudv$$

for some function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$ such that

(a) $f_{X,Y}(u, v) \geq 0$ for all $u, v \in \mathbb{R}$

(b) $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u, v)dudv = 1$.

Then X and Y are jointly continuous *and $f_{X,Y}$ is their* joint density function.

Of course,

$$f_{X,Y}(x, y) = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

For a single continuous random variable X, it turns out that the probability that it lies in some nice set $A \in \mathbb{R}$ (see Part A Integration to see what we mean by "nice", but note that any set you can think of or write down will be!) can be obtained by integrating its density over A:

$$\mathbb{P}(X \in A) = \int_A f_X(x)dx.$$

Likewise, for nice sets $B \subseteq \mathbb{R}^2$ we obtain the probability that the pair (X, Y) lies in B by integrating the joint density over the set B:

$$\mathbb{P}((X, Y) \in B) = \int \int_{(x,y) \in B} f_{X,Y}(x, y)dxdy.$$

We will show here that this works for rectangular regions B.

Theorem 7.22. *For a pair of jointly continuous random variables X and Y, we have*

$$\mathbb{P}\left(a < X \leq b, c < Y \leq d\right) = \int_c^d \int_a^b f_{X,Y}(x,y)dxdy,$$

for $a < b$ and $c < d$.

Proof. We have

$$\begin{aligned}
&\mathbb{P}\left(a < X \leq b, c < Y \leq d\right) \\
&= \mathbb{P}\left(X \leq b, Y \leq d\right) - \mathbb{P}\left(X \leq a, Y \leq d\right) + \mathbb{P}\left(X \leq a, Y \leq c\right) - \mathbb{P}\left(X \leq b, Y \leq c\right) \\
&= F_{X,Y}(b,d) - F_{X,Y}(a,d) + F_{X,Y}(a,c) - F_{X,Y}(b,c) \\
&= \int_c^d \int_a^b f_{X,Y}(x,y)dxdy.
\end{aligned}$$

□

Definition 7.23. *If X and Y have joint density $f_{X,Y}$ then the* marginal densities *of X and Y are*

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dy$$

and

$$f_Y(y) = \int_{-\infty}^{\infty} f_{X,Y}(x,y)dx$$

respectively.

These definitions generalise straightforwardly to the case of n random variables, $X_1, X_2, \ldots, X_n$.

Example 7.24. *Let*

$$f_{X,Y}(x,y) = \begin{cases} \frac{1}{2}(x+y) & \text{for } 0 \leq x \leq 1,\ 1 \leq y \leq 2, \\ 0 & \text{otherwise.} \end{cases}$$

Check that $f_{X,Y}(x,y)$ is a joint density. What is $\mathbb{P}\left(X \leq \frac{1}{2}, Y \geq \frac{3}{2}\right)$? What are the marginal densities? What is $\mathbb{P}\left(X \geq \frac{1}{2}\right)$?

Solution. Clearly, $f_{X,Y}(x,y) \geq 0$ for all $x, y \in \mathbb{R}$. We have

$$\begin{aligned}
\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X,Y}(x,y)dxdy &= \int_1^2 \int_0^1 \frac{1}{2}(x+y)dxdy \\
&= \int_1^2 \left[\frac{1}{4}x^2 + \frac{1}{2}xy\right]_0^1 dy \\
&= \int_1^2 \left(\frac{1}{4} + \frac{1}{2}y\right) dy \\
&= \left[\frac{1}{4}y + \frac{1}{4}y^2\right]_1^2 \\
&= 1.
\end{aligned}$$

We have

$$\begin{aligned}\mathbb{P}\left(X \leq \frac{1}{2}, Y \geq \frac{3}{2}\right) &= \int_{3/2}^{2}\int_{0}^{1/2} \frac{1}{2}(x+y)dxdy \\ &= \int_{3/2}^{2}\left[\frac{1}{4}x^2 + \frac{1}{2}xy\right]_0^{1/2} dy \\ &= \int_{3/2}^{2}\left(\frac{1}{16} + \frac{1}{4}y\right)dy \\ &= \left[\frac{1}{16}y + \frac{1}{8}y^2\right]_{3/2}^{2} \\ &= \frac{1}{4}.\end{aligned}$$

Integrating out y we get

$$f_X(x) = \int_1^2 \frac{1}{2}(x+y)dy = \frac{1}{2}x + \frac{3}{4}$$

for $x \in [0,1]$, and integrating out x we get

$$f_Y(y) = \int_0^1 \frac{1}{2}(x+y)dx = \frac{1}{4} + \frac{1}{2}y$$

for $y \in [1,2]$. Using the marginal density of X,

$$\mathbb{P}\left(X \geq \frac{1}{2}\right) = \int_{\frac{1}{2}}^{1}\left(\frac{1}{2}x + \frac{3}{4}\right)dx = \frac{9}{16}.$$

□

Definition 7.25. *Jointly continuous random variables X and Y with joint density $f_{X,Y}$ are* independent *if*

$$f_{X,Y}(x,y) = f_X(x)f_Y(y)$$

for all $x, y \in \mathbb{R}$. Likewise, jointly continuous random variables $X_1, X_2, \ldots, X_n$ with joint density $f_{X_1,X_2,\ldots,X_n}$ are independent *if*

$$f_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = f_{X_1}(x_1)f_{X_2}(x_2)\ldots f_{X_n}(x_n)$$

for all $x_1, x_2, \ldots, x_n \in \mathbb{R}$.

Note that if X and Y are independent then it follows easily that

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)$$

for all $x, y \in \mathbb{R}$.

Example 7.26. *Consider the set-up of Example 7.24. Since there exist x and y such that*

$$\frac{1}{2}(x+y) \neq \left(\frac{1}{2}x + \frac{3}{4}\right)\left(\frac{1}{4} + \frac{1}{2}y\right),$$

X and Y are not independent.

7.5.1 Expectation

We define the expectation of a function h of a pair of jointly continuous random variables in the obvious way:

$$\mathbb{E}\left[h(X,Y)\right] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(x,y) f_{X,Y}(x,y)dxdy.$$

In particular, the *covariance* of X and Y is

$$\mathrm{cov}\left(X,Y\right) = \mathbb{E}\left[(X-\mathbb{E}\left[X\right])(Y-\mathbb{E}\left[Y\right])\right] = \mathbb{E}\left[XY\right] - \mathbb{E}\left[X\right]\mathbb{E}\left[Y\right]$$

(exercise: check the second equality).

Exercise 7.27. *Check that*

$$\mathbb{E}\left[X+Y\right] = \mathbb{E}\left[X\right] + \mathbb{E}\left[Y\right]$$

and

$$\mathrm{var}\left(X+Y\right) = \mathrm{var}\left(X\right) + \mathrm{var}\left(Y\right) + 2\mathrm{cov}\left(X,Y\right).$$

Remark 7.28. *We have now shown that the rules for calculating expectations (and derived quantities such as variances and covariances) of continuous random variables are* exactly the same *as for discrete random variables. This isn't a coincidence! We can make a more general definition of expectation which covers both cases (and more besides) but in order to do so needs a proper theory of Integration, which you will see in Part A. For the moment, we just observe that the fact that everything works the same way for continuous random variables as it does for discrete ones has the consequence that the things we did in Chapter 6 applies for continuous random variables too. In particular, it makes perfect sense to think about a random sample from a continuous distribution, and you will deal with these extensively in Prelims Statistics next term.*

Example 7.29. *Let $-1 < \rho < 1$. The standard bivariate normal distribution has joint density*

$$f_{X,Y}(x,y) = \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)$$

for $x, y \in \mathbb{R}$. What are the marginal distributions of X and Y? Find the covariance of X and Y.

Proof. We have

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right)dy \\ &= \int_{-\infty}^{\infty} \frac{1}{2\pi\sqrt{1-\rho^2}}\exp\left(-\frac{1}{2(1-\rho^2)}[(y-\rho x)^2 + x^2(1-\rho^2)]\right)dy \\ &= \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi(1-\rho^2)}}\exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right)dy. \end{aligned}$$

But the integrand is now the density of a normal random variable with mean ρx and variance $1-\rho^2$. So it integrates to 1 and we are left with

$$f_X(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2}.$$

So $X \sim \mathrm{N}(0,1)$ and, by symmetry, the same is true for Y. Notice that X and Y are only independent if $\rho = 0$.

Since X and Y both have mean 0, we only need to calculate $\mathbb{E}[XY]$. We can use a similar trick:

$$\begin{aligned}\mathbb{E}[XY] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \frac{xy}{2\pi\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}(x^2 - 2\rho xy + y^2)\right) dydx \\ &= \int_{-\infty}^{\infty} \frac{x}{\sqrt{2\pi}} e^{-x^2/2} \int_{-\infty}^{\infty} \frac{y}{\sqrt{2\pi(1-\rho^2)}} \exp\left(-\frac{(y-\rho x)^2}{2(1-\rho^2)}\right) dydx.\end{aligned}$$

The inner integral now gives us the mean of a $\mathrm{N}(\rho x, 1-\rho^2)$ random variable, which is ρx. So we get

$$\operatorname{cov}(X,Y) = \int_{-\infty}^{\infty} \frac{\rho x^2}{\sqrt{2\pi}} e^{-x^2/2} dx = \rho\mathbb{E}\left[X^2\right] = \rho,$$

since $\mathbb{E}\left[X^2\right] = 1$. □

This yields the interesting conclusion that standard bivariate normal random variables X and Y are independent if and only if their covariance is 0. This is a nice property of normal random variables which is *not true* for general random variables, as we have already observed in the discrete case.

Common discrete distributions

Distribution	Probability mass function	Mean	Variance	Generating function
Uniform $\mathrm{U}\{1,2,\ldots,n\}$, $n \in \mathbb{N}$	$\mathbb{P}(X=k) = \frac{1}{n}, 1 \le k \le n$	$\frac{n+1}{2}$	$\frac{n^2-1}{12}$	$G_X(s) = \frac{s-s^{n+1}}{n(1-s)}$
Bernoulli $\mathrm{Ber}(p)$, $p \in [0,1]$	$\mathbb{P}(X=1) = p$, $\mathbb{P}(X=0) = 1-p$	p	$p(1-p)$	$G_X(s) = 1-p+ps$
Binomial $\mathrm{Bin}(n,p)$, $n \in \mathbb{N}, p \in [0,1]$	$\mathbb{P}(X=k) = \binom{n}{k} p^k (1-p)^{n-k}$, $k = 0,1,\ldots,n$	np	$np(1-p)$	$G_X(s) = (1-p+ps)^n$
Poisson $\mathrm{Po}(\lambda)$, $\lambda \ge 0$	$\mathbb{P}(X=k) = \frac{\lambda^k}{k!}\mathrm{e}^{-\lambda}$, $k = 0,1,2,\ldots$	λ	λ	$G_X(s) = \mathrm{e}^{\lambda(s-1)}$
Geometric $\mathrm{Geom}(p)$, $p \in [0,1]$	$\mathbb{P}(X=k) = (1-p)^{k-1}p$, $k = 1,2,\ldots$	$\frac{1}{p}$	$\frac{1-p}{p^2}$	$G_X(s) = \frac{ps}{1-(1-p)s}$
Alternative geometric, $p \in [0,1]$	$\mathbb{P}(X=k) = (1-p)^k p$, $k = 0,1,\ldots$	$\frac{1-p}{p}$	$\frac{1-p}{p^2}$	$G_X(s) = \frac{p}{1-(1-p)s}$
Negative binomial $\mathrm{NegBin}(k,p)$, $k \in \mathbb{N}, p \in [0,1]$	$\mathbb{P}(X=n) = \binom{n-1}{k-1}(1-p)^{n-k}p^k$, $n = k, k+1,\ldots$	$\frac{k}{p}$	$\frac{k(1-p)}{p^2}$	$G_X(s) = \left(\frac{ps}{1-(1-p)s}\right)^k$

Common continuous distributions

Distribution	Probability density function	Cumulative distribution function	Mean	Variance
Uniform $\mathrm{U}[a,b]$, $a<b$	$f_X(x)=\dfrac{1}{b-a},\ a\le x\le b$	$F_X(x)=\dfrac{x-a}{b-a},\ a\le x\le b$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$
Exponential $\mathrm{Exp}(\lambda)$, $\lambda\ge 0$	$f_X(x)=\lambda \mathrm{e}^{-\lambda x},\ x\ge 0$	$F_X(x)=1-\mathrm{e}^{-\lambda x},\ x>0$	$\dfrac{1}{\lambda}$	$\dfrac{1}{\lambda^2}$
Gamma $\mathrm{Gamma}(\alpha,\lambda)$, $\alpha>0,\lambda\ge 0$	$f_X(x)=\dfrac{\lambda^\alpha}{\Gamma(\alpha)}x^{\alpha-1}\mathrm{e}^{-\lambda x},\ x\ge 0$		$\dfrac{\alpha}{\lambda}$	$\dfrac{\alpha}{\lambda^2}$
Normal $\mathrm{N}(\mu,\sigma^2)$, $\mu\in\mathbb{R}$, $\sigma^2\ge 0$	$f_X(x)=\dfrac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x\in\mathbb{R}$	$F_X(x)=\Phi\left(\dfrac{x-\mu}{\sigma}\right)$	μ	σ^2
Standard Normal $\mathrm{N}(0,1)$	$f_X(x)=\dfrac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}},\ x\in\mathbb{R}$	$F_X(x)=\Phi(x)=\displaystyle\int_{-\infty}^{x}\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}\,du$	0	1
Beta $\mathrm{Beta}(\alpha,\beta)$	$f_X(x)=\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1},\ x\in[0,1]$		$\frac{\alpha}{\alpha+\beta}$	$\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$